6

ISSUES IN MACHINE AIDED TRANSLATION: USING PARALLEL CORPORA AS A RESOURCE

Mona Parakh

Introduction

In the1950s Chomsky turned linguistics away from empiricism and towards rationalism in a short span of time. He underpinned the corpus as a source of evidence in linguistic studies. However, in the 1980s there was a resurgence of corpus-based work in linguistics as the computer gradually became the mainstay of corpus linguistics. The availability of the computerised corpus and the wider availability of institutional and private computing facilities do seem to have provided a spur to the revival of corpus linguistics (McEnery and Wilson, 1996).

Nowadays the term "corpus" has become synonymous with the term "machine-readable corpus". A corpus, simply defined, is a large body of text. Corpora may exist in machine readable form or in their natural state as written texts or recorded speech, but increasingly the term "corpus" is used to refer to the machine readable variety (McEnery and Wilson, 1993).

The corpora are not just a collection of texts from various sources, but there is a lot more, from selection of corpora to the development of tools and inferring knowledge from the corpora. Corpus linguistics, like all linguistics, is concerned primarily with the description and explanation of the nature, structure and use of language (Arora et al.). Corpus linguistics methods have found way into theoretical and descriptive branches of translation. This has made it possible to draw generalisations about the nature and the function of language as well as the frequent yet unconventional recurrences of language patterns observable over a large collection of texts or corpora. As is commonly known, corpus linguistics is becoming increasingly important for translation studies (Baker, 1995).

In dealing with the issues of translation, this work makes use of the English-Hindi parallel corpus of automatically aligned sentences. These aligned sentences have been generated by aligning texts from the English and Hindi editions of the 'India Today' Corpus using an automatic sentence aligner (see section 3).

Parallel Corpora

Corpora can be of various types depending on its use. This work makes use of the parallel corpus. A variety of corpora-types are covered by the term parallel corpora, but in general it refers to texts that are

translations of each other. A parallel corpus is not immediately user-friendly. For the corpus to be useful it is necessary to identify which sentences in the sub-corpora are translations of each other, and which words are translations of each other. A corpus which shows these identifications is known as an aligned corpus as it makes an explicit link between the elements which are mutual translations of each other (McEnery and Wilson, 1996).

Parallel corpora may be aligned at different levels such as text alignment, sentence alignment, chunk alignment, word alignment. Parallel corpus aligned at the level of texts does not provide any useful insight into the nature of the translation; hence this study has been based on a corpus of aligned sentences.

Typical applications of parallel corpora include, construction of lexicons and grammars, Bilingual/Multi-lingual lexicography, Translator training, Machine learning, etc. Parallel corpora are also of use to those involved in cross-linguistic research and comparative linguistic research.

Basic Algorithms for Automatic Alignment of Sentences

Sentence alignment is the problem of making explicit the relations that exist between the sentences of two texts that are known to be mutual translations (Simard and Plamondon, 1996). In general, what a sentence aligner does is that it takes two aligned texts as input and gives aligned sentences as output.

There are a few existing algorithms which do good alignment. One such algorithm is the Gale and Church Algorithm (1993). The Gale and Church Algorithm is based on a very simple statistical model of character lengths. The model makes use of the fact that longer sentences in a source language are likely to be translated into longer sentences in the target language, and that shorter sentences are likely to be translated into shorter sentences. A probabilistic score is assigned to each pair of proposed sentence pairs, based on the ratio of lengths of the two sentences (in characters) and the variance of this ratio. This probabilistic score is used in a dynamic programming framework in order to find the maximum likelihood alignment of sentences.

For some languages the task of alignment has proved to be difficult. In such languages it is difficult to use the statistical analysis of sentence lengths to do the alignment. Further, there are sizeable additions and deletions that can occur in either the source or target languages, and more so when the languages are far apart. Besides, a lot of sentences align many-to-many and this makes the task even more difficult.

While some of the efficient algorithms ignore word identities and consider only the sentence lengths, Akshar Bharati et al. (2002) describe an algorithm for aligning sentences with their translations in a bilingual corpus, using lexical information of the languages. For a sentence in the source language text, the algorithm picks the most likely translation from the target language text using lexical information and certain heuristics.

The data used to test the algorithm came from a weekly news magazine "India-Today". The magazine is released in two languages. The source language is English and the target language into which it is later translated is Hindi.

The various resources that the algorithm uses are English-Hindi Bilingual Dictionary, Monolingual Hindi Dictionary, Morphological Analyser and Local-Word Grouper. The algorithm does not do any statistical analysis using sentence lengths. The algorithm is language independent and it also aids in detecting addition and deletion of text in translations.

Translation Issues

In the following section, some issues in translation have been dealt with using the English-Hindi parallel corpus of automatically aligned sentences. These sentences from the English and Hindi editions of the 'India Today' Corpus have been aligned using the algorithm for automatic sentence alignment by Akshar Bharati et al. (2002), explained in the previous section.

One of the common issues in text translation is the conflict between the naturalness of the translation and its faithfulness towards the source text. So, often in an attempt to keep the translation faithful to the source, a compromise has to be made on the fluency and readability of the target or translation text, and vice versa. Interestingly, this conflict provides insightful perspectives into the nature of the translation. Other issues that develop as a result of the conflict are the differences between the source-text and its translation which can be accounted for by the differences in the nature and structure of the languages in question.

Nida (1959) discusses some basic underlying principles of translation which state that "no translation in a receptor language can be the exact equivalent of the model in the source language. That is to say, all types of translation involve (1) loss of information, (2) addition of information, and/or (3) skewing of information."
For machine aided translation, these are vital issues given that, while translating texts, human translators rely to a great extent on various linguistic and non-linguistic resources. Non-linguistic sources or background knowledge includes common sense knowledge, world-knowledge, domain specific knowledge and contextual references. On the other hand, machines do not have similar ontological or common sense resources. Though there have been efforts in the direction of building lexical networks and ontologies, such knowledge based resources fail to match up with the knowledge of a human translator.

With the help of some examples it is possible to illustrate how the differences between the source text and its translation can be accounted for, in terms of certain linguistic devices used by human translators. These are problems that cannot be easily handled by machines.

Inversion

Inversion of 'chunks' across English-Hindi aligned sentences can be accounted for in terms of the difference in their word order. English has a word order of the SVO (Subject-Verb-Object) type whereas Hindi generally follows the SOV (Subject-Object-Verb) convention, though it has a relatively free word order.

Indian languages have a relatively free word order. Many of the constituents of a simple sentence can occur in any order without affecting the gross meaning of the sentence; what is affected is perhaps the emphasis (Akshar Bharati et al., 2000).

Let us take the following examples where the sentences (a) and (b) are automatically aligned and are mutual translations of each other.

1. (a) [This gigantic migratory fish]1 [has been sought out]2 [in Gujarat since ancient times]3 [for its liver oil]4

1. (b) [isa vishaalkaaya pravaasii machalii ke]1 [jigara ke tela ke lie]4 [gujaraata meM praachiina kaala se hii]3 [isakii kaaphii maaMga rahii hei]2 *

As is clear from the above example, the order and arrangement of word 'chunks' (marked by numbered brackets) varies across English and Hindi, due to the difference in their word order. The position of the chunks 2 and 4 in the English sentence 1.(a) has been interchanged in the Hindi translation 1.(b). As mentioned earlier, since Hindi has a relatively free word order, such an inversion of chunks does not affect the overall meaning of the sentence. However, inversion of chunks in English would change the gross meaning of the sentence.

Omission

At times information that exists in the source text may be missing in its translation. The translator may on purpose omit certain information, in order to maintain the readability, fluency and the naturalness of the translation.

Take for example the following aligned sentences:

2. (a) There are [many others among] Marandi's 26 ministers, including 10 belonging to alliance parties
2. (b) maraaMDii ke maMtrimaMDala meM 26 maMtrii heiM, jinameiM se 10 to gaThajoDa melM shaamila paartiyoM ke heiM.
3. (a) More confident the second time around, he again approached Aamir and [got] the endorsement he was seeking.
3. (b) agalii baara jyaadaa aatmavishvaasa ke saatha ve phir aamir ke paas pahuuzche aur unakii sviikruti chaahii.

In the above examples, the information 'many others among' in 2.(a) is missing in the corresponding Hindi translation 2.(b). Since such constructions are not natural to Hindi, the translator has chosen to exclude that information rather than impose an unnatural construction. Besides, by leaving out that construction, as such no information is lost.

However, in the third example, the omission of the word 'got' in the Hindi sentence 3. (b), leads to a loss of information. While the word 'got' in the English sentence 3. (a) suggests that the task of getting the endorsement was accomplished, in the Hindi translation 3. (b) the omission of the word 'got' causes a loss of information, which implies that the task of getting the endorsement was not achieved. The omission in this case changes the meaning of the translation.

Insertion

The translator may at times, insert certain information in the target text that does not exist in the source. This is done depending on whether there is any previous reference of the additional information within the context of the text. At times the translator may insert additional information on the basis of his own "subjective" knowledge. Take the following sentences for example.

4. (a) There are several [others] who have constantly been pressurising the chief minister to post "nice and cooperative" officials in their departments.

4. (b) ese kaii [maMtrii] heiM, jo apane vibhaagoM meM "acche our sahayogii" adhikaariyoM kii niyukti ke lie mukhyamaMtrii para dabaava Daalte rahe heiM.

5. (a) A little later,[he] and two other colleagues were summoned by the patrol party.

5. (b) Kuch dera baada [ganaaii] our unake do anya sahakarmiyoM ko gashtii dala ne bulaa liyaa.

Nothing in the English sentences tells us that the 'others' in 4. (a) and the 'he' in 5. (a) refer to 'maMtrii' and 'ganaaii' respectively. But the translator has added that information on the basis of his own knowledge or by reference to the context. While such addition of information is normal for human translators, it poses a significant challenge for machine aided translation.

Substitution

Translators often "take the liberty" to substitute words in the target text with words that are somewhat related in sense, but are not exact equivalents of words in the source text. Translators generally take recourse of near equivalents when a good and exact equivalent in the target language is unavailable or non existent. On the other hand, it may also be possible, that due to the translator's insufficient knowledge of either the source or target language he/she may be incapable of finding a suitable corresponding term for a word in the source text.

6. (a) Officials [pointed out] that as chairman he would be heading a tribunal

6. (b) adhikaariyoM ka [tarka thaa] ki adhyaksh ke ruupa meM ve eka paMchaaT ke mukhiyaa hoMge

As seen in the above example, the translator has substituted the expression 'pointed out' in 6. (a) with the expression 'tarka thaa' in the Hindi translation 6. (b). The expression 'tarka thaa' which means 'reasoned out' is not an equivalent but rather a substitution for the expression 'pointed out' in 6. (a).

Other Linguistic Issues

There are other linguistic issues that come up due to the devices commonly employed by human translators in the course of translating from one language to another. Certain language structures give rise to linguistic issues such as paraphrasing, translating tense, aspect and modality of verbs, translating relative clauses and translating prepositions. In the context of translation these linguistic issues are related, in that, they are problems pertaining to substitution. They are also challenging issues from the perspective of machine aided translation.

Paraphrasing [One word against group of words]

When a translator fails to find an equivalent for a term in the source language, generally he/she substitutes the term with a closely related one in the target language. This has been dealt with under the issue of substitution (see section 4.4). Another way by which translators address this issue is by employing descriptive phrases i.e., providing an explanation or definition of the source language term, in the target language.

7. (a) it is expected to be welcomed by parents who feel subjects like history are [overdone] and a burden on children.

7. (b) Ummiida hei, ve maataa – pitaa isakaa svaagata kareMge jinakaa maananaa hei ki itihaasa jeise viSayoM kaa [jaruurata se jyaadaa samaavesha hotaa hei] our ve baccoM para bojha heiM.
8. (a) It is not the number which is [exceptional] but the BRES' almost totally communal syllabus.

8. (b) lekina [dhyaana dene laayaka baata] saMkhyaa nahiiM balki biiaariies ka laga bhaga puurii taraha saaMpradaayika paaThyakrama hei.

As seen in the above examples, the single word expressions 'overdone' in 7. (a) and 'exceptional' in 8. (a) do not have exact equivalents available in Hindi and hence, the translator has resorted to paraphrasing them in 7. (b) and 8. (b) respectively, thereby representing single terms with groups of words.

TAM [Tense Aspect Modality]

Saeed (1997) while explaining the concepts of tense, aspect and modality of a verb, states that tense allows a speaker to locate a situation relative to some reference point in time, most likely the time of speaking. Aspects have to do, not with the location of an event in time, but with its temporal distribution or contour. Modality is a cover term for devices which allow speakers to express varying degrees of commitment to, or belief in, a proposition.

Translation of Tense, Aspect and Modality of verbs is another challenge often encountered during machine translation. It is difficult to establish a one to one mapping of Tense, Aspect and Modality across languages.

As in the example below, the verb 'can't believe' in the English sentence 9. (a) is in the simple present tense, whereas its translation in 9. (b) is in the present continuous tense. If in this example, a simple present tense form is imposed on the Hindi sentence in 9. (b), then the resulting sentence would be unnatural in accordance to the sentence structure of Hindi.

9. (a) Jamie Whitby and Katherine Katkit [can't believe] their eyes.

9. (b) jemii vhitbi our keithariina keiTkiTa ko apanii aazkhoM para [bharosaa nahiiM ho rahaa hei].

Given that the TAM of a verb in a particular language can be mapped variously into different languages, a translator may take recourse of his/her proficiency in the languages concerned to settle on the appropriate alternative. However, for a machine, such a task would be quite complicated.

Relative Pronouns

Languages vary in terms of the structures they use to code information and so where one language may use a relative structure to code certain information, another language, due to lack of a similar structure, may code the information differently. This means, that a relative clause in the source language need not necessarily get translated as a relative clause in the target language. The following example illustrates this point.

10. (a) Subjects like history are a burden on children [who] should focus on "contemporary" and "job-oriented" science and mathematics.

10. (b) itihaasa jeise viSaya baccoM para bojha heiM, [jabaki unheM] gaNita our vignyaan jeise 'saamayika' aur 'rojgaaronmukhii' viSaya paDhaae jaane caahie.

In the English sentence 10. (a) the relative pronoun 'who' is used as a post modifier of the word 'children'. However, in the Hindi translation 10. (b) the relative pronoun 'who' has been translated as a pronoun meaning 'they.'

Preposition

Another issue in translation is seen in terms of the differences while translating prepositions. The purpose that prepositions serve in English is served through post-position markers in Hindi. Besides, the post-position markers in North Indian Languages (such as Hindi) and case endings in South Indian Languages play a key role in specifying semantic relationships between verbs and their arguments (Akshar Bharati et al., 2000). Given the above, it is difficult to establish a one to one mapping of prepositions in English to post-positions in Hindi, as is evident in the following examples.

11. (a) The actors sport the lean, hungry look-not [of]1 an Amitabh Bachchan [in]2 Deewar but perhaps [of]3 a Balraj Sahni [in]4 Do Bigha Zameen.

11. (b) isake kalaakaara dubale-patale, bhuukhe-naMge lagate hein, diivaara [ke] 2 amitaabha baccana [jeise] 1 nahiiM balki do biighaa jamiina [ke] 4 balaraaja saahanii [jeise] 3.

The preposition 'of' in sentence 11.(a) is translated as 'jaise' (meaning 'like') in the Hindi sentence 11. (b). While the preposition 'in' is translated as 'ke' (meaning 'of') in the Hindi sentence. Since prepositions in English can be translated variously into Hindi, the task of choosing the appropriate equivalent from a number of alternatives is a difficult task for the machine.

Conclusion

This paper has outlined the nature and use of parallel corpora, with particular reference to aligned corpus and the methods for automatic sentence alignment. The paper has concentrated specifically upon the use of parallel corpora to identify issues that arise in machine aided translation.

Note

* The Hindi example sentences in this paper are in accordance with the Roman Notation for Devanagari followed by Akshar Bharati et al. (2000).

References

Akshar Bharati, V.Chaitanya,R. Sangal.2000. *Natural Language Processing: A Paninian Perspective*. New Delhi: Prentice-Hall of India.

Akshar Bharati, et al. 2002 . An Algorithm for Aligning Sentences in Bilingual Corpora Using Lexical Information. *Proceedings of ICON-2002.* Bombay

Baker, M. 1995. Corpora in Translation Studies: An Overview and Some Suggestions for Future Research. *Target* 7(2). 223-243.

Gale, W. & K. Church. 1993. A Program for Aligning Sentences in Bilingual Corpora. *Computational Linguistics* 19(1).91-1023.

Mcenery, T. and A. Wilson.1996. *Corpus Linguistics.* Edinburgh: Edinburgh University Press.

Nida, E.1959. Principles of Translation as Exemplified by Bible Translating. R. Brower (ed.), On *Translation.* Cambridge, Massachusetts : Harvard University Press

Saeed, J. 1997) *Semantics.* Oxford: Blackwell

Simard, Michel and Pierre Plamondon 1996. Bilingual Sentence Alignment: Balancing Robustness and Accuracy. *Proceedings of the First Conference of the Association for Machine Translation in the Americas (AMTA-96).* 135-144