A Study Into The Design Of A Soft-Computing Based Unsupervised Translator That Deals With Ambiguities

Shatadru Sengupta, Bipasha Biswas, Apratim Mitra, Sk Shahnawaj,
Department of Computer Applications, Haldia Institute of Technology, Haldia 721657, West Bengal, India

Abstract:
A key area of work, and indeed of challenges, in Machine Translation (MT) is dealing with ambiguity. While in MT we aim to reduce the gap between human perception and machine solution, we have repeatedly failed to show perfect results regarding this, and we continue to failing getting human like translations from our machines. This conceptual paper reviews a few of the commonly faced problems in ambiguity resolution when translating from English to Bengali and vice versa, and, on the basis of this, aims to propose a soft-computing based method in designing an English to Bengali and vice versa translator which may yield better results in the future. The implementation and hence the corresponding analytical studies thereafter are left as future work.
Keywords: NLP, Machine Translation, MT, Natural Language Processing, Unsupervised removal of ambiguity, Genetic Algorithm, Fuzzy Logic and NLP

List of Tables:

List of Figures:

1. Introduction:

Machine Translation (a part of NLP) is a part of Artificial Intelligence [4], having a wide range of applications from multilingual device manuals to enhanced customer service [8]. However, since human speech is plagued with ambiguities, meaning that a word or a sentence, and indeed sometimes a phrase, can mean more than one different things. It does not beggar saying that this creates miscommunication among humans.

In MT, these ambiguities account for some of the most prolific problems that can create havoc with the ultimate result. For example, imagine a multilingual product manual; - the manual is first written in, say, English by a group of content writers or documentation officers and then, we may use an MT system to produce versions of this manual in other languages like, say, Bengali. Since the original manual has been written in natural English, if there is bound to be some ambiguity hidden in the original text, which will raise its ugly head in the translation and lead to

a different meaning than what is intended. This error may then lead to a disaster in using the product.

The main types of ambiguities we deal with in this paper, which can be exhibited in a given piece of input text, are explained below [6][7]:

1. Syntactic ambiguity – when the sentence can have two different meanings (eg. The bus hit the kerb in the red area: was the kerb in the red area or was the bus hit in some red area on its body?)
2. Semantic: when either a single word has two or more meanings (*polysemy:* foot of a hill vs foot of a man) or two different words have the same pronunciations leading to an ambiguity in hearing (*homonymy:* dear and deer)
3. Lexical ambiguity – when a single word can have more than one meanings (bank – to deposit money or riverbank or rely, pen – writing instrument or enclosure for animals)
4. Structural ambiguity – when the ordering of words in a sentence alters the meaning of the sentence (e.g. I saw a white girl with my binoculars. Ambiguity: girl with my binoculars or girl seen through my binoculars. However, With my binoculars, I saw a white girl – this has a clear meaning that the white girl was seen through my binoculars.)

In effect, it is very difficult to translate from one language to another and at the same time eliminate the ambiguities imbibed in the sentences. Truly, the nature and manner and the chances of arising of any ambiguity will completely depend upon what is being said. Hence we need some computing tool(s) so that ambiguity resolution can be carried out without much intervention from human beings. In section 2, we will see a few examples of ambiguities that may arise when we try to translate from English to Bengali sentences and vice versa. We shall also point out what type of ambiguity we are looking at.

2. Some findings regarding English-Bengali and Bengali -English Translations

Let us see some examples of translation as observed from Google Translate:

A. English to Bengali	***B. Bengali to English***
1. Bank situated at the riverside: *nadir tire* ***(bank = river bank, human: a financial bank which is situated by the river)*** ***→ semantic and lexical ambiguity***	1. Nadir tire: *On the river bank (**bank = river bank, human: river bank)*** ***→ no ambiguity at all***
2. Bank on the riverside: *nadir tire (**bank = river bank, human: a financial bank which is situated by the river)*** ***→ semantic and lexical ambiguity***	2. Nadir tire abasthito tir: *the banks of the river(**bank = river bank, human: wrong sentence)*** - attempted reverse translation w.r.t. A.3 ***→ syntactic ambiguity***

3. Bank situated on the riverside: *nadir tire abasthito tir ((**bank = river bank, human: a financial bank which is situated by the river)*** ***→ semantic, lexical as well as syntactic ambiguity***	3. Jibikar jonyo nadir upor bhorosa: *Relying on the river for livelihood* ***(humans can say either relying, or banking – but Google translate does not recognise banking as a synonym for relying)*** ***→no ambiguities at all***
4. Bank on the river: *nadir tire (**bank = river bank, human: river bank, or* **depend on the river*)*** ***→ semantic, lexical as well as structural ambiguity***	4. Dabanale tar murgir ghor pure gechhe: *the fire burnt his chicken coop.* ***(humans use both hen pen and chicken coop, but Google translate cannot recognise pen as coop) → no ambiguities at all***
5. Bank on the river for livelihood: *jibikar jonyo nadir tire ((**bank = river bank, human: depend on the river)*** ***→lexical ambiguity***	
6. His pen of hens was burnt in the wildfire: *dabanale tar murgir kalam pure gechhe* ***(pen = writing instrument, human: an enclosure to keep hens in) →lexical ambiguity***	
7. His sheep pen was sold off: *tar bherar kalam bikri hoye gechhe* ***(pen = writing instrument, human: an enclosure to keep hens in) →lexical ambiguity***	
8. Animal pen: *poshu kalam* ***(pen = writing instrument, human: an enclosure to keep hens in) → lexical ambiguity***	

Table 1: Ambiguities exhibited by online Google Translate

Table 1: We can see from the Table 1 that in the two words, pen and bank, have been translated in quite a different manner to how a human being would translate, and as indicated in the table itself, exhibits all of syntactic, semantic and lexical ambiguity while using an established online translation platform, namely Google Translate, even in simple sentences.

While translations A.1 to A.8 all exhibit ambiguity, so does reverse translation B.1. B.2, B.3, and B.4 do not have any ambiguities.Thus we see that dealing with ambiguity is crucial in making artificial speech more human-like.

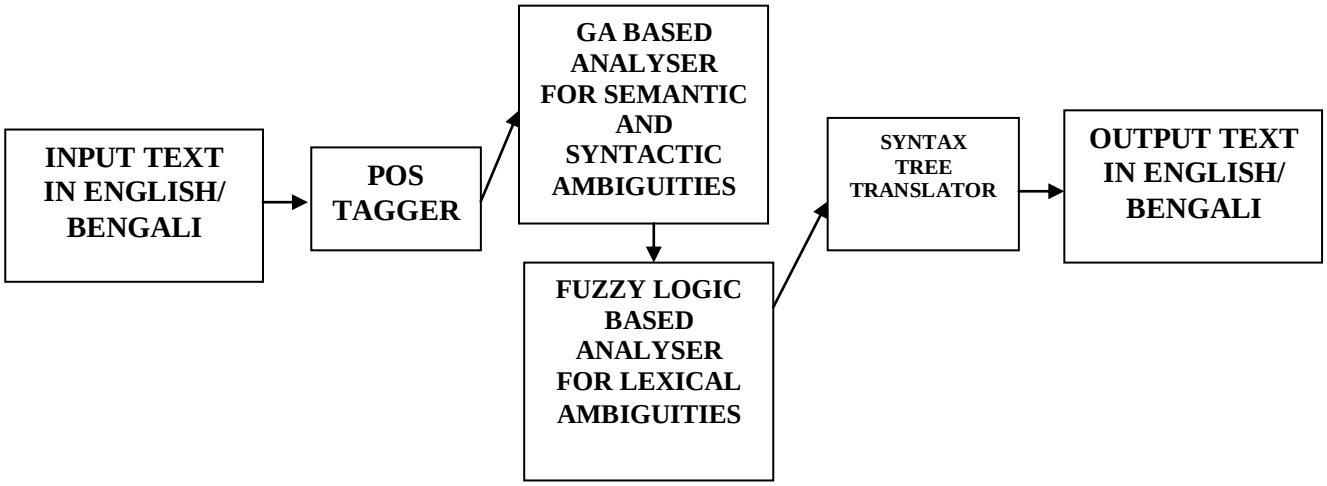
3. Dealing with Ambiguity through Unsupervised Learning

Dealing with lexical, semantic and syntactic ambiguities are generally solved through soft computing techniques and form a core part of Machine Translation, and hence of NLP. The nature of data that we will deal with here is obviously very complex. It is observed from Section 2 that a human being takes the context into account while dealing with translation. This means s/he remembers the information about the conversation or sentence whenever s/he analyses it. Needless to mention, a machine translator will have to have the same capacity in order to come up with a translation that matches human translation.

Since natural language related ambiguities can arise at any point of time, and may be thitherto unseen or not experienced, we would like to adapt to an unsupervised approach towards ambiguity resolution. Thus, Genetic Algorithms and Fuzzy Logic [5] may be considered as two representatives of such unsupervised methods. In this regard, Michael Berk [1] is of opinion that GPs can perform well on data that exhibit complex nature. Again, since we are dealing with an Indic language, we may consider that in [2], Mamulkar and Nandanwar say that amongst unsupervised techniques, GA gives better results while translating from Marathi.

Fuzzy logic also is a forefront technique when it comes to dealing with uncertainty, as is the case with ambiguities. Gupta, Jain and Joshi [3] mention that a fuzzy based solving system works in close resemblance with the human problem solving characteristics, exactly the approach that we need here. They also praise on to say that from unorganised information, Fuzzy systems can approximate reasonably good solutions.Thus it is clear that use of GA and Fuzzy based systems may be successfully used to perform human like translations. The question now is how to bring these two technologies together in one single unit of translation. We propose a hybrid system as follow:

FIGURE 1: THE STEP BY STEP AMBIGUITY CHECKER



Explanation of Figure 1:

In Figure 1 we present what we call the "Step By Step Ambiguity Checker". The input text is first taken through a "POS Tagger and Syntax Tree Translator". Those of us who have not dealt with MT before will appreciate the fact that this is a stage where the input sentences are first broken down into the corresponding words and these are then tagged as being one of the parts of either Bengali or English speech, as the case may be (POS Tagging).

Next, according to predefined human notions of word sense, the genetic algorithm based subsystem will try to remove the semantic and syntactic ambiguities available in the input text. This will involve analysing the context and of the sentences and changing them into a less ambiguous form which is easily relatively easy to translate. The semantic and syntactic ambiguities will form a huge huddle to cross as the entire input text is going to have to be analysed as a whole in order to get a context sensitive output from this stage.The Fuzzy logic based analyser will then work specifically in solving lexical ambiguities. The accuracy of output of this stage will depend on the accuracy of the GA based analyser.

Finally, from the reviewed, analysed and unambiguous intermediate version of the input text, we can proceed with the syntax tree translator to convert from English/Bengali syntax tree to the corresponding Bengali/English syntax tree and ultimately form the target output text.

4. Conclusion

Translation from Bengali to English and vice versa will be a challenging sequence of work given that the alphabet of Bengali and English are different in size and most of the sentences and rules are not compatible. Of particular problem will be the fact that several types semantic, syntactic and lexical ambiguities will mar the system from producing the same translation that a human being will produce. Our study into this matter, in light of a small part of the literature available on this issue, seems to reveal that a mixture of unsupervised soft-computing strategies may be the solution to this hindrance. We propose that before doing syntax tree translation, GA and Fuzzy Logic based ambiguity analysis for firstly ambiguities in the semantics and then the syntax followed by the lexical ambiguities maybe solved from the POS tagged syntax tree. After this ambiguity resolution, the revised syntax tree can be translated into the target language. In future, we will have to implement, analyse, test and report each part of this translation system.

5. References

[1] Berk, Michael. "Genetic Algorithms for Natural Language Processing" Towards Data Science.
https://towardsdatascience.com/genetic-algorithms-for-natural-language-processingb055aa7c14e9?gi=ee90b1dffb0a
[2] Mamulkar, Kalyani; Nandanwar, Lokesh. "Marathi Word Sense Disambiguation using Genetic Algorithm – A Review". IJACEN, ISSN 2320-2106, Volume 3, Issue 5, May 2015

[3] Gupta, Charu; Jain, Amita; Joshi, Nisheeth. "Fuzy Logic in Natural Language Processing – A Closer View". Procedia Computer Science 132 (2018), Volume 132, 2018, pages 1375 – 1384, Elsevier Publications.
[4] Patterson, Dan W. "Artificial Intelligence and Expert Systems". 2004 edition. ISBN-81-203-0777-1. Page 227 – 269.
[5] Sivanandam, S. N. Deepa, S.N. "Principles of Soft Computing". Second Edition. Wiley India Private Limited. ISBN- 978-81-265-2741-0. Chapters 7 and 15.
[6] Ambiguity: Definitions and Examples. https://literaryterms.net/ambiguity/
[7] Ambiguity. https://en.wikipedia.org/wiki/Ambiguity
[8] https://www.bloomreach.com/en/blog/2019/09/natural-language-processing.html#