**CSME 206A – Natural Language & Speech Processing** **Spring Semester**

# Lecture 16: Statistical machine Translation-I

*Lecturer: K.R. Chowdhary* *: Professor of CS*



## 16.1 Introduction

Proficiency in languages was traditionally a hallmark of a learned person. Although the social standing of this human skill has declined in the modern age of science and machines, translation between human languages remains crucially important, and MT is perhaps the most substantial way in which computers could aid human-human communication. Moreover, the ability of computers to translate between human languages remains a consummate test of machine intelligence: Correct translation requires not only the ability to analyze and generate sentences in human languages but also a human-like understanding of world knowledge and context, despite the ambiguities of languages.

Machine translation (MT) is the automatic translation from one natural language into another using computers. The field of machine translation (MT) aims to develop computer systems that are able to translate text or speech without human intervention. However, present translation technology has not been able to deliver fully automated high-quality translations. Typical solutions to improving the quality of the translations supplied by an MT system require manual post-editing. This serial process prevents the MT system from taking advantage of the knowledge of the human translator, and the human translator cannot take advantage of the adaptive ability of the MT system.

Machine-translation systems work with vocabulary databases and algorithms that implement grammatical rules. The translation engine typically looks up known words first in its database. The system then applies its grammatical rules to the source text and, using various algorithms, tries to figure out meanings of other words using clues from their position and the grammatical construction. The engine then uses the target language's grammatical rules to construct translated sentences.

An alternative way to take advantage of the existing MT technologies is to use them in collaboration with human translators within a Computer-Assisted Translation (CAT) or interactive framework. Historically, CAT and MT have been considered different but close technologies and more so for one of the most popular CAT technologies, namely, *translation memories*. Interactivity in CAT has been explored for a long time. Systems have been designed to interact with human translators in order to solve different types of ambiguities, like, *lexical*, *syntactic*, or *semantic*. Other interaction strategies have been considered for updating user dictionaries or for searching through dictionaries.

An important contribution done in the the field of CAT technology is that it entailed an interesting focus shift in which interaction is directly aimed at the production of the target text, rather than at the disambiguation of the source text, as in earlier interactive systems. The idea is to embed data-driven MT techniques within the interactive translation environment, with the aim to combine the best of both paradigms: CAT, in which the human translator ensures high-quality output, and MT, in which the machine ensures a significant gain in productivity.

Following the above ideas, the innovative embedding above consists in using a complete MT system to produce full target sentence hypotheses, or portions thereof, which can be accepted or amended by a human translator. Each correct text segment is then used by the MT system as additional information to achieve further, i.e., improved suggestions. More specifically, in each iteration, a prefix of the target sentence is somehow fixed by the human translator and, in the next iteration, the system predicts a best (or $n$-best) translation suffix(es) to complete this prefix. This process can be called as Interactive-Predictive Machine Translation (IPMT). This approach introduces two important requirements: i.e., the models have to provide adequate completions and, this has to happen efficiently.

In this chapter, we shall study about *discriminative models* that do not use only local associations to produce the target sentence. But instead, these models generate every lexical unit in the target sentence by considering the context of the entire source sentence. We term this approach of selecting target language lexical items by considering features of the entire source sentence as Global Lexical Selection (GLS). Although, in our discussion, we consider only the lexical features of the source sentence, but, the GLS approach can incorporate syntactic information as well as determine the lexical items of a target sentence. These lexical items are then *reordered* using a language model.

Statistical machine translation (SMT) is an approach to MT that is characterized by the use of machine learning methods. The SMT draws from many fundamental research areas in computer science, so some knowledge of automata theory, formal languages, search, and data structures are beneficial to understand SMT. Familiarity with statistical theory and mathematical optimization techniques used in machine learning are also be helpful, but SMT focuses on the main ideas and intuitions behind the mathematics rather than full mathematical rigor, majority of them have been covered in the first chapter.

The interest in SMT (Statistical Machine Translation) is attributed to many factors, some of them are:

1. The growth of the Internet has increased consumption of translated text by those who disseminate information in multiple languages, e.g., multilingual governments and news agencies, and by the companies operating in the global marketplace. The Internet enables them to easily publish information in multiple languages. In addition, due to this widespread dissemination, SMT researchers now have access to bilingual news text, and other data mined from the Internet. These data are also the basic resource in SMT research. Because these data are product of day-to-day human activities, they are constantly growing. Multilingual governments interested in dissemination, such as the European Union, have increased MT research funding to further their domestic policy interests.
2. The other type of consumers of translation are those interested in the assimilation of information not in their native language. These include intelligence agencies, researchers, and casual Internet users. The Internet has made such information much more readily accessible, and increasing demand from these users helps drive popular interest in MT.
3. Fast and cheap computing hardware has enabled applications that depend on large data and billions of statistics. Advances in processor speed, random access memory size, secondary storage, and grid computing have all helped to enable SMT.
4. The development of automatic translation metrics – although controversial – has enabled rapid iterative development of MT systems and fostered competition between research groups.
5. Several projects have focused on the development of freely available SMT toolkits. Many are open-source.

Most of the earlier research on statistical machine translation (SMT) is based on word-alignment algorithms, which provide local associations between source words and target words. The source-to-target word-alignments are sometimes augmented with target-to-source word alignments in order to improve the precision

of these local associations. Further, the word-level alignments are extended to phrase-level alignments in order to increase the extent of local associations. The phrasal associations compile some amount of (local) lexical reordering of the target words – those permitted by the size of the phrase. Most of the state-of-the-art machine translation systems use these phrase-level associations in conjunction with a target language model to produce the target sentence. There is relatively little emphasis on (global) lexical reordering other than the local reordering permitted within the phrasal alignments. A few exceptions are the hierarchical (possibly syntax-based) transduction models.

Unlike most approaches for machine translation which allow training of only limited number of features, *Direct Translation Models* (DTM) relies on training all its parameters using the *Maximum Entropy Model*. This model states that the probability distribution that best represents the current state of knowledge about a system is the one having largest entropy. The maximum entropy is the one that makes the fewest assumptions about the true distribution of data. However, it still uses only local associations and mild context information (previous word and next word) during translation.

## 16.2 Subproblems SMT

Following these core ideas we discussed in the beginning of this chapter, there are four problems that we must solve in order to build a functioning SMT system.

1. First, we must describe the series of steps that transform a source sentence into a target sentence. We can think of this as creating a story about how a human translator might perform this task. This story is called a *translational equivalence model*, or more simply a model. All of the translational equivalence models that we consider, derive from concepts from automata and language theory.
2. Next, we want to enable our model to make good choices when faced with a decision to resolve some ambiguity. We need to develop a parameterization of the model that will enable us to assign a score to every possible source and target sentence pair that our model might consider.

   Taken together, translational equivalence modeling and parameterization are often combined under the rubric of modeling.
3. The parameterization defines a set of statistics called parameters used to score the model, but we need to associate values to these parameters. This is called *parameter estimation*, and it is based on machine learning methods.

## 16.3 Machine Translation of Indian Languages

The problem of machine translation can be viewed as consisting of two sub-problems: (a) *lexical selection*, where appropriate target language lexical items are chosen for each source language lexical item, and (b) *lexical reordering*, where the chosen target language lexical items are rearranged to produce a meaningful target language string. Both these problems are extremely challenging, especially in the case of translation from English to Indian languages and vice-versa. In addition to the lack of a large sentence aligned corpora, lexical choice for Indian languages is difficult because of (1) morphological richness of Indian languages and (2) agreement between features of syntactic related words in Indian languages. Lexical reordering is also a major issue between English and Indian languages as the degree of *word-reordering* is extremely high. This is in contrast to relatively local lexical reordering between, e.g., English and French. However, Indian languages are relatively *free-word order languages* where the grammatical role of the content words is largely

determined by its case markers and not by their positions in the sentence. Hence, predicting appropriate *function-words* associated with the content words is perhaps more crucial than ordering the words in the target language sentence.

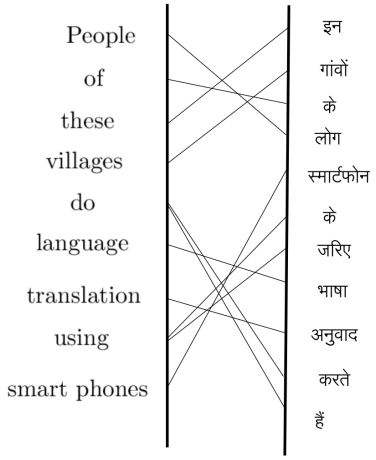


Figure 16.1: Sample English sentence and its Hindi translation

### 16.3.1 Bag of Words Model

The *simple bag of words* (BOWs) model is a simplest model. Given a source sentence, each of the target words are chosen by considering the lexical features of the entire source sentence. The selected target language words are then permuted in various ways and each permutation is ranked using a target language model and the best permutation is chosen as the target language translation. The size of the search space of these permutations can be set by a parameter called the *permutation window*. This model does not allow long distance reordering of target words unless a very large permutation window is chosen – a computationally very expensive proposition.

In the bag of local-word-groups (lwgs) model, we explore the prediction of lwgs in the target language. The lwgs are groups of words having one content word and the associated function words. This model is important for two reasons. Besides selecting appropriate words in the target sentence, it also predicts the association between the content words and the function words. This association conveys the grammatical roles of the content words. To obtain a well-formed target sentence, these local word groups are ordered appropriately using a target language model.

### 16.3.2 Factored Sequential Lexical Choice Model

To solve the problem of BOWs model, a simple model, called *sequential lexical choice model* (SLCM) is used. This model generates words in an order which is faithful to the order of words in the source sentence. Now, the number of permutations that need to be examined to obtain the best target language strings is much less when compared to the bag-of-words model. This model is expected to give good results for language pairs such as English-French for which only local word order variations exist between sentences.

A *factored sequential lexical choice model* (FSLCM) handles the morphological richness, lexical choice, and association of functions words to appropriate content words and issues of data sparsity. In contrast to the bag of lwgs, where content words and associated function words are predicted together, in this model they

are predicted independently. The morphological features and the function words of the content words are treated as factors of the content words. The factors are predicted independently and associated to the source positions using sequential lexical choice model, thereby associating them with each other. The factors are then combined using a morphological generator to obtain the target language local word groups, which are then ordered appropriately.