Translation Challenges and Universal Networking Language

Baljeet Kaur Dhindsa
Assistant Professor
Guru Gobind Singh College for Women
Sector 26, Chandigarh, India

Dharam Veer Sharma
Associate Professor
Department of Computer Science
Punjabi University, Patiala, India

ABSTRACT
This paper presents challenges being faced in designing automatic translation software. There are many approaches to automatic translation like Direct, Rule based, Transfer based, Statistical based and Interlingua. A brief description of all the approaches and their advantages and drawbacks are discussed. Universal Networking Language (UNL) based on Interlingua approach can be used especially for translation among multiple languages because it requires knowledge of UNL and of the language which user wants UNL to support. User can then get translated text in any of the languages supported by UNL without even being oblivious to any other language. It is less expensive approach also. This paper also gives brief introduction to UNL and how it can overcome many of the challenges in translation.

Keywords
Machine Translation, Machine Translation Approaches, Interlingua, Challenges in Machine Translation, Universal Networking Language, Overcoming challenges in Machine Translation using UNL.

1. INTRODUCTION
Language is one facet that discerns man from all other species on the earth, because it is one asset that no other creature possesses. On one hand, it allows a man to express himself and on the other, it facilitates him in gaining knowledge. Knowledge is assumed to be a service to mankind as it results in new ideas, new information and it helps in acquiring more knowledge also. There are numerous languages in the world and India is linguistically diverse country as according to 2001 census, there are 22 constitutionally scheduled languages, 100 mother tongues and approximately 1000 documented languages and dialects. India is a country which symbolizes unity in diversity. According to Marc Twain, India is the place where human languages originated, history evolved, legends born, and diverse cultures developed. Multiple languages give language freedom to its people, but it also results in perplexity among masses due to usage of different languages in different parts of the same country. Freedom of language is the birthright of every individual, so to overcome with the problems due to multiple languages, few simple measures can be taken. One of the best possible solutions, in the current era of technological revolution can be providing tools which can translate data into the language of the user. The word *Translation* means converting text, written/spoken, from one language to another. For communication among people with different linguistic usage, there arises a need to translate written/spoken language of one group of people for understanding by the other group who are trying to communicate with one another. For effective communication, two people need to understand each other's language appropriately either by using the same language or by using a mediator who can translate language used by one person to another person's language. Communication between sender and receiver group(s) is incomplete if the receiver does not understand the information from the sender effectively in the context in which it has been communicated.

Human translators are doing translation among group of people having different language of communication since centuries. Although human translators are impeccable in their performance, yet scarcity of translation specialists is a problem for the growing translation markets around the world. However, with the advancement in technology, this process has become automated to some extent. Translation nowadays is not restricted to communication of one's ideas only, but it is required by people for expansion of their business also. Automatic Machine Translation is conversion of text from one language to another by using a machine. Many automatic translation procedures are in use these days. Some of these automatic machine translators need human intervention for effective translation while some are capable of doing it automatically. Due to increase in demand of translation and that too at faster speed as well as in multiple languages simultaneously, machine translation has become a necessity in today's fast paced global interactions. Although the study and development of automatic language translators has been going on since 1947 [18], yet it is becoming an increasingly important topic for researchers due to increased requirement of communication in the era of globalization and use of numerous languages worldwide.

2. CHALLENGES IN TRANSLATION
Translating written text can pose several challenges which are illustrated as follows:

2.1 Language Structure
Languages can be differentiated in the way they order words in sentences; some languages follow SOV (Subject-Object-Verb) structure like Hindi, Punjabi, Hungarian, Turkish, Japanese etc.; while some follow SVO (Subject-Verb-Object) like English, Malay, Germanic, Thai etc. Some languages follow VSO structure like Irish, Arabic, Hebrew etc. [7] and very few languages follow VOS structure like Palauan, Tzotzil [12]. Translation may become difficult when a sentence in source/target language follows one type of structure and in target/source language follows another.

2.2 Translate or Transliterate
This is another challenge in translation field that how to decide whether a word in source language is to be translated or transliterated, where transliteration means representing word given in source language into a word in target language, in which the character set of the target language only need to be used, keeping the pronunciation of the original word intact. Some words in a sentence are required to be translated, whereas some words do not need to be translated but rather transliterated like name of countries, states, people, places, organizations etc. In some cases it becomes difficult to make such a differentiation among words. For example, a simple

rule in transliteration is, names of country, names of states, company names etc. are transliterated, but "India", although it is name of country needs to be translated to "भारत" [Bharat] in Hindi language, country name "China" becomes "चीन" [Cheen] in Hindi language. Another issue is that in some multiword names, some word might need translation, while some might need transliteration e.g."Mount Everest" should be written as "ऐवरेस्ट पर्वत" [Everest Parvat] in Hindi language, so here "Everest" needs to be transliterated and "Mount" needs to be translated.

2.3 Tokenization

First step in any translation system is tokenization, which is analysis of the sentence so as to divide it into tokens or individual words [6]. In most of the languages tokenization is easy as white space or punctuation is used as word boundary, but there are some languages where it is difficult to find tokens to differentiate words like Chinese, Japanese, where words are not delimited by common delimiters: white space or punctuations [5]. Urdu is another such language. In such cases additional algorithms are required first to tokenize sentence and then to translate, which makes the translation procedure more difficult.

2.4 Homograph

Some languages support homographs, which are words having same spellings but entirely different meanings like English, Chinese, Japanese [5]. To choose correct meaning of such words one needs to know context information of that word, which is difficult to judge sometimes, especially by a machine. For example, "Heart attack" means "दिल का दौरा" [Dil ka Daura] in Hindi; here attack means "दौरा" [Daura], whereas in the sentence "cats attack rats", attack means "हमला" [Hamala].

2.5 Speaker's Attitude

Attitude of the speaker can change the meaning of a word and in return that of the sentence. In a written text, it is difficult to judge the speaker's intentions and therefore might result in incorrect translation. Sometimes a positive sentence, but said in derisive manner can mean something entirely different. For example: "Do you really think he is brilliant?" can mean, the person is asking whether someone is brilliant or not, but if this sentence is said in sarcastic manner then it can mean that the speaker is doubting person's brilliance.

2.6 Languages with multiple scripts

Some languages support multiple scripts like Punjabi (Gurmukhi, Shahmukhi), Kashmiri (Devnagri, Perso-Arabic), Japanese (Kanji, Hiragana, Katakana) etc. Multiple scripts make translation all the more difficult, because one needs to form a lot of rules for translation for all types of scripts.

2.7 Language Writing Systems

Languages all over the world can be categorized into three categories based on the writing systems they use, which are Alphabetic systems, Syllabic and Logographic languages [13]. Alphabetic system uses a small set of letters or symbols called alphabet representing phoneme of language as in English, Hindi, Punjabi language etc.; Syllabic uses a set of written symbols called syllabary representing syllables, which are used to constitute words in that language as in Korean, Japanese language; Logographic uses logograms, where a logogram is a single character representing a grammatical word in that language as in Chinese language. Many other European languages use logograms like & (and), @, $ etc. along with alphabets. Translation becomes difficult for languages using multiple writing systems. Logographs pose difficult challenges when needs to be translated especially.

2.8 Ambiguity due to Grammatical Variations

Every language has some common features: Gender, Number, Prepositions etc., but some variations are there in many languages. Gender: Some languages allow two genders (Masculine/Feminine) e.g. Hindi, French, Spanish etc., while some languages allow three, neutral being the third, e.g. Sanskrit, English etc. Moreover, gender affects Pronouns and verbs [14]. Number: In some languages, number system is of two types (Singular/Plural) e.g. English, Hindi etc., while a few languages have three types, dual being the third, e.g. Sanskrit, Arabic and to some extent Hebrew. In such languages, plural is used for clusters with more than two items. Number affects noun, pronoun, verb and adjective in a sentence [14]. Similarly, modifiers are placed before noun in some languages, while in others these are placed after nouns. Such variations in languages pose challenges while translating text, especially when source language follows one type of features and target language another.

Languages might encounter some more challenges due to more variations in languages like right-to-left syntax (Urdu, Hebrew etc.) or left-to-right syntax (Hindi, English, French etc.) of sentences or due to similar script but different meanings of the words like "angel" means "sting" in Dutch, "fishing pole" in German. Similar variations in many languages can be found. Similarly it is difficult to translate idioms and phrases, more so because idioms are language/area dependent many times like, "Doodh ka jala chaach bhi phoonk-phoonk kar peeta hai" is a very famous idiom in Hindi, which translates to "Once bitten twice shy" in English language. Many idioms depend on the language they are made in and it is difficult to translate in another language using a program, unless these are hard coded.

3. MACHINE TRANSLATION APPROACHES

History of machine translation goes back to 1949, when Warren Weaver combined Statistical method and information theory to propose solution to the problem of ambiguity in languages [18]. Since then many approaches came into existence. Major approaches which are popular and are successfully being used in developing translators are:

3.1 Dictionary Based/ Direct Translation

Direct Translation System [8, 16] is the oldest approach in machine translation history and is still being used, in which translation unit is a word. It is a two steps process; in step I, for every word in source text, its corresponding meaning is found in the target language using a large bilingual dictionary of source language and target language and in step II some simple grammatical adjustments are done e.g. on word order or morphological, which are based on grammatical rules of target language. In this case source language text is not analyzed structurally beyond morphology and therefore has following drawbacks:

- It is bilingual and unidirectional.
- It is not suitable for the sentences where the meaning is concept based e.g. idioms, phrases, slogans etc. "A pen is mightier than a sword", "it's a cakewalk" are

examples of idioms, which cannot be translated word by word.

- User might not get grammatically correct text in target language always. It can be useful for only those people who are familiar in the target language to some extent and therefore can understand meaning of grammatically incorrect sentences somehow.

3.2 Rule Based

In Rule Based Systems [1, 17], set of rules are made for every type of legitimate sentence in the source language and another set of rules are made for sentences in the target language. These rules are made considering the grammatical rules of the corresponding languages. Whenever a sentence is given to be translated, it is compared with the set of rules for that language and then bilingual dictionary is consulted for finding the meaning of the words in the target language. The requisite rule in the target language is found and the sentence is displayed in the target language. Rules can be used at every stage of translation i.e. syntactic, semantic and contextual processing stage. Although it has advantages like it can be extended to multiple languages and is useful for dynamic languages where vocabulary keeps on adding and new rules can be added accordingly, but it has following drawbacks:

- It is a very exhaustive approach as it requires entire linguistic acquaintance of both the languages by the developer.
- It requires complete knowledge of syntactically correct sentences, which must be reproduced for each and every entity definition because it does not adhere to the syntactic generality.
- It is domain specific also.

3.3 Transfer Based

In Transfer Based Systems [2, 20], text written in source language is converted into an intermediate language and then this intermediate language is used to convert text in target language. In fact, the intermediate language of source text is used to convert text into intermediate language of target text and then finally it is converted into target language. This intermediate language is dependent on source language and target language to some extent. It is mainly used for translating data in compatible languages. Its major drawbacks are:

- It is suitable for bilingual translation only.
- It performs well only for the compatible source and target languages.

3.4 Statistical Based Systems

This is the most common approach in use these days. Statistical Based systems [10, 19] requires bilingual corpus for each language pair. It is based on the assumption that every sentence in the target language is a translation of some sentence in the source language. Statistical technique utilizes the concept of probability theory given in statistics. Let SL represents text in source language and TL represents text in target language. Given the text TL, text SL is found from which translator generated TL. To reach at this decision, probability values are assigned to every pair of SL and TL. Probability values of (SL,TL) pairs would be low or nil if TL does not represent SL properly and probability value will be higher for those (SL,TL) pairs where TL is translated version of SL. SL is chosen based on maximum probability value of the pair (SL,TL). It has advantages like it is language independent, fast and less expensive provided that a good corpus of bilingual texts is available and also developer is not required to have knowledge of source and target languages. In spite of that, it has few drawbacks also:

- Its success is largely dependent on the availability of good quality bilingual corpus, which is a difficult task and moreover it requires a lot of space to store that.
- Its success is also dependent on estimating the probability values of the text, which is a difficult task because it requires a lot of training.
- One sentence can result in more than one translation, then which of the translated text will get what probability value, needs to be resolved.

3.5 Interlingua Based

Interlingua based systems [4, 9, 11, 15] works on a sentence, where sentence in source language is converted into language independent intermediate language, which is then used to convert text into target language. This technique is similar to that used in Transfer Based Systems, but in Interlingua based systems intermediate language is language independent. It is more beneficial because this feature enables this approach to support multilingual translation. Universal Networking Language (UNL) and Lexical Conceptual Structures (LCS) are two very popular intermediate languages. Its major advantages are that it supports multilingual translation, its easier to add support to more languages by adding analysis and synthesis component of new languages and it is economical as compared to transfer based system. Its drawbacks are:

- It is more time-consuming as a lot of processing time is consumed in the 'double-transfer'.
- The biggest disadvantage of using this approach is that it requires set of primitives to be defined. These primitives allow cross language mapping. Defining primitives is a difficult process.

4. UNL FOR TRANSLATION CHALLENGES

All the five approaches discussed above have its advantages and disadvantages, but if support to multiple languages is required for the translation then there are two approaches: Statistical based and Interlingua based. Statistical based requires a big corpus in all the languages requiring translation, which requires a lot of manual effort, time, money and storage space. On the other hand, Interlingua approach is comparatively economical because it requires 2n converters to be developed, where 'n' is number of languages requiring translation and after that all these 'n' languages can do translation among themselves. Moreover it is easier to extend this system for a new language, which requires addition of analysis and synthesis pair for the new language. These points make Interlingua approach the strongest, if one wants support for multiple languages in translation.

UNL [22] is one of the most promising intermediate languages, which can be used to translate one language into multiple languages without writing code for set of all these languages. UNL is an artificial language. It is registered in the name of United Nations to ensure unlimited accessibility for all. A lot of research is going on UNL all over the world. UNL is an Interlingua, proposed by United Nations University, Tokyo, Japan, to access, transfer and process information available in natural language. It can translate one language to every interface language in its system. It is better than natural language in the sense that expressions in natural language can be ambiguous, but not in UNL. UNL

expressions are always unambiguous. Currently it supports English, Japanese and Chinese, although a lot of work has also been done for the languages like French, Arabic and also some Indian languages like Punjabi, Bangla and Malayalam etc. UNL requires development of two core software: EnConverter and DeConverter. EnConverter is a language independent parser, which provides framework for morphological, syntactic and semantic analysis synchronously and converts source language sentences into UNL expression. DeConverter is a language independent generator, which provides a framework for syntactic and morphological generation. It converts UNL expression into target language. This will result in UNL giving support to source language.

A sentence in UNL is expressed as UNL expression, consisting of a set of binary relation(s), where each binary relation represents relationship between tokens of the source text. Every token in source text is converted into Universal Word (UW), where UW is corresponding word in UNL representing the token in source text. UNL is the intermediate language and every language is converted into UNL so that it can be translated into target languages by the corresponding DeConverter software. Every UW represents a defined concept and it can also include constraint list(s) to represent ambiguous words, which can remove ambiguity. It can also include attribute(s) to describe speaker's attitude. For example, following is a sentence in English:

"This computer translates from English to Hindi."

Its corresponding UNL expression[21]:

agt(translate(icl>do).@entry, computer(icl>machine))
mod(computer(icl>machine),this)
src(translate(icl>do).@entry,english(icl>language))
gol(translate(icl>do).@entry, hindi(icl>language))

Here, four binary relations are required to represent the given sentence: agt, mod, src, gol. There are 56 already defined relations in UNL and user can use any one of these only. In the given UNL expression, first binary relation is:

agt(translate(icl>do).@entry, computer(icl>machine))

'agt' is relation as described above, 'translate' is first UW (UW1), '(icl>do)' is constraint list and '@entry' is attribute. Similarly, 'computer' is second UW (UW2), '(icl>machine)' is its constraint list.

'agt' relation represents that UW2 is agent/initiator of UW1.

'mod' relation represents that UW2 is modifier/attribute of UW1.

'src' relation represents that UW2 is source of UW1.

'gol' relation represents that UW2 is goal of UW1.

In UNL, UWs are represented by English language, so here English is used as intermediate language.

Constraint list consist of 'icl' which means 'a kind of', so translate is a kind of 'do' means action, computer is a kind of machine, English and Hindi are kind of languages. Constraint list can include multiple constraints to remove ambiguity. Here, '@entry' simply represents that the corresponding UW is entry point or major concept of the given sentence. Similarly @present, @past, @future attributes can be used to represent tense of source text, @interrogative to represent query sentences. There are 94 attributes in all which can be used by developer to describe subjectivity of sentences.

UNL can meet many of the challenges in translation, discussed above, which are described as follows.

- **Language Structure:** A sentence in UNL is represented as a collection of binary relations as discussed above, so it does not follow any natural language structure like SOV, SVO etc. Developer needs to understand purpose of 56 relations and 94 attributes in order to be familiar with UNL.
- **Translate/Transliterate:** In UNL, a token that needs to be transliterated has 'iof' (instance of) relation in its constraint list, which specifies that the corresponding token needs to be transliterated and not translated.
- **Tokenization:** Every UW represents a token in UNL, so developer can get tokens as UW in UNL expression. Second step of tokenization is to differentiate tokens as verb, noun, adjective etc. which can be achieved by carefully analyzing constraint list and attributes of every UW. For example, in the above example, translate is used as verb, which can be known from its constraint list 'icl>do' which means translate is a kind of action. Similarly English and Hindi are representing languages here, which can be known by analyzing their constraint list 'icl>language', which means it is a kind of language.
- **Homograph:** In UNL, ambiguity can be removed between tokens having different meanings with the help of constraint list and/or attributes only as explained in the above point. For example, English(icl>people) means 'अंग्रेज' (Angrez i.e. English man/woman) , but English(icl>language) means 'अंग्रेजीभाषा' (Angrezibhasha i.e. English language).
- **Speaker's Attitude:** In UNL, attributes are assigned to capture speaker's feelings, judgment, point of view and attitude. This concept is already explained above.

Some of the challenges mentioned above do not hold here like language writing system, ambiguity due to grammatical variations, left-to-right/right-to-left syntax etc. Idioms/phrases require hard coding and that cannot be handled in UNL like most of the other approaches.

5. CONCLUSION

Language translation is in use since centuries as human translators were doing it earlier and are still doing it, but its need has increased further and that also in multiple languages. Globalization and fast paced information generation and exchange are the major contributors to its increase. Automatic language translators can be helpful in enhancing business opportunities as business organizations need to interact in various languages in various countries and they can reach out to a wide range of people by delivering information of their business in native languages of their prospective customers. It has the advantage of saving time and money as it is faster and has high throughput as compared to that by human translator [3]. Interlingua translators like UNL can be more beneficial in such circumstances as these are more economical. People in different countries are contributing to it in a major way, but machine translation has still not reached to the level of human translators as far as quality of translation is concerned.

6. REFERENCES

[1] A. Kupsc, "Two Approaches to Aspect Assignment in an English-Polish Machine Translation System," in proceedings of 7th International EAMT Workshop on Machine Translation and other Language Technology Tools, Hungary, pp. 17-24, 2003.

[2] A. Lavie, S. Vogel, L. Levin, E. Peterson, K. Probst, A. F. Llitjos, R. Reynolds, J. Carbonell, and R. Cohen, **"**Experiments with a Hindi-to-English transfer-based MT system under a miserly data scenario," inACM Transactions on Asian Language Information Processing (TALIP), Vol. 2, Issue 2, pp. 143-163, 2003.

[3] A. Yanishevsky, "The Emerging role of Machine Translation," [Online] Available: http://www.promt.ru/press/pdf/promt_for_multilingual.pdf [Accessed: Oct 26, 2015].

[4] B. Dorr, "Parameterization of the Interlingua in Machine Translation," in the proceedings of the 14th Conference on Computational Linguistics, (COLING '92), Vol. 2, pp. 624-630, 1992.

[5] C. Olinsky and A. W. Black, "Non-Standard Word and Homograph Resolution for Asian Language Text Analysis," in ICSLP-2000, 2000.

[6] E. Rich, and K.Knight, "Artificial Intelligence", Second Edition,Tata McGraw-Hill, pp: 377-424, 1991.

[7] G. A. Broadwell, "It ain't necessary S(V)O: Two Kinds of VSO Languages," in Proceedings of the LFG05 Conference, 2005. [Online] Available: http://web.stanford.edu/group/cslipublications/cslipublications/LFG/10/lfg05broadwell.pdf.[Accessed: Dec 10, 2015].

[8] G. S. Josan, G. S. Lehal, "A Punjabi to Hindi Machine Translation System," in the proceedings of the 22nd International Conference on Computational Linguistics, (COLING '08), pp. 157-160, 2008.

[9] J. Lee, and S. Seneff, "Interlingua-Based Translation for Language Learning Systems," in Automatic Speech Recognition and Understanding, IEEE Workshop, pp. 133-138, 2005.

[10] P. F. Brown, J. Cocke, S. A. D. Pietra, V. J. D. Pietra, F. Jelinek, J. D. Lafferty, R. L. Mercer, and P. S. Roossin,"A Statistical Approach to Machine Translation," in Computational Linguistics, Vol. 16, Number 2, pp. 79-85, 1990.

[11] R. K. Mohanty, M. K. Prasad, L. Narayanaswamy, and P. Bhattacharyya, "Semantically Relatable Sequences in the Context of Interlingua Based Machine Translation," in the proceedings of the 5th International Conference on Natural Language Processing, Hyderabad, (ICON 2007), pp. 1-8, 2007.

[12] S. Chung, "Properties of VOS Languages", 2005, pp. 685688.[Online]Available:http://people.ucsc.edu/~schung/chung_syncom.pdf [Accessed: Dec 10, 2015]

[13] S. Karimi, F. Scholer, and A. Turpin, "Machine Transliteration Survey," ACM Computing Survey, Vol. 43(3), pp. 1-46, 2011.

[14] S. B. Singh, "English-Hindi Translation Grammar," Prabhat Prakashan, pp. 82-186, 2010.

[15] T. Mitamura, E. H. Nyberg, and J. G. Carbonell, "An Efficient Interlingua Translation System for Multi-lingual Document Production," in the proceedings of Machine Translation Summit III, Washington D.C., pp. 105-117, 1991.

[16] V. Goyal, G. S. Lehal, "Web Based Hindi to Punjabi Machine Translation System," in Journal of Emerging Technologies in Web Intelligence, Vol. 2, No. 2, pp. 148-151, 2010.

[17] V. Kommaluri, "Rule-based Machine Translation System using Indian Logic for Discourse Texts," Report, C-DAC Mumbai, Mar 2007. [Online] Available: http://tdil.mit.gov.in/april-jan-2008/8.16_Rule-basedMachine%20Translation.pdf.[Accessed Nov 12, 2010].

[18] W. J. Hutchins, "Machine Translation over fifty years", in Histoire, Epistemologie, Langage, Tome XXII, fasc. 1, pp.7-31, 2001.

[19] W. Kraaij, J. Y. Nie, and M. Simard, "Embedding Web-Based Statistical Translation Models in Cross-Language Information Retrieval,"in Association for Computational Linguistics, pp. 381-419, 2003.

[20] W. Winiwarter, "Machine Translation Using Corpus-based Acquisition of Transfer Rules", in the proceedings of the International Conference on Digital Information Management, (ICDIM '07), pp. 345-350, 2007.

[21] Y. Grishkyan, "Upon Comparison of Some Online UNL Modules," in Proceeding of CSIT2009, 2009.

[22] Universal Networking Digital Language Foundation, [Online] Available:http://www.undl.org. [Accessed: Dec 15, 2015]