An Evaluation of Output Quality of Machine Translation Program

Mitra Shahahbi
MA. Student
University of Wolverhampton
Stafford Street
Wolverhampton WV1 1NA
United Kingdom
Shahabi_mitra@yahoo.com

ABSTRACT

This article reports an exploratory evaluation of the output quality of two prevalent English-Persian Machine Translation programs. The purpose of the research is to find out which program produces relatively better output, and what major linguistic bottlenecks MT programs will encounter in their processing of texts. Criteria were established in light of structural theories to solve the MT output from the perspective of Accuracy and Intelligibility. For each program, the mean score it obtained for its output and the rate of correctness of its translation of the testing points were calculated. An analysis of the mean score and the rate of correctness of each program generated the following findings about the output quality of these two programs: 1) Padideh Translator produces the best output. 2) The major linguistic bottlenecks in English-Persian MT programs occur in the areas of morphology, complex sentences, syntactic ambiguity and semantic analysis, generation of Persian, and long sentences.

KEYWORDS

Machine Translation, Accuracy, Intelligibility

1. INTRODUCTION

MT technology is very important in the future of business. More and more business is being done on the Internet. People from every country are starting to consider the World Wide Web a mall from which they can buy anything they need [3]. Although English is spoken widely across world; it is the 4th most spoken language, and is by far the most extensively used language to communicate science, propagate technology and do business, not all potential users have access to this language This leaves many potential customers who do not understand the English-only websites on the internet. MT helps the business adapt to the customers.

The economic necessity of finding a cheaper solution to international exchange has resulted in continuing technological progress in terms of translation tools designed to automate and computerize the translation of natural language texts or to use computers as an aid to translation [2].

Although MT has some disadvantages, we will be able to use MT for cheaper and faster translation in near future. The present, relatively poor quality of translation yield by the computer: where total grammatical, semantic/associative meanings and pragmatic adequacy are concerned, it can lead the native speaker to reject the text on the grounds that it is strange, awkward, and even nonsensical [1] when reading. But there are also good reasons why we use machines to translate our texts. The primary reasons for using machine translation are speed, cost savings, and availability. John Hutchins [8] summarizes the reasons for using computers in translation as follows and insists any one of these may justify MT or computer aids:

- Too much translation for humans
- Technical materials too boring for humans
- Greater consistency required
- Need results more quickly
- Not everything needs to be top quality
- Reduce costs

MT prompts researchers to ask whether it is possible that we have MT systems that can produce translation that is as good as human translation but faster and cheaper. What programs produce relatively better translation? And what difficulties are most MT programs confronted with? This article intends to probe into these questions and report an exploratory evaluation of the output quality of two prevalent English-Persian MT programs, namely *Pars Translator* and *Padideh Translator.*

Student Research Workshop, RANLP 2009 - Borovets, Bulgaria, pages 71–75

1.1. Machine Translation Evaluation (MTE)1

The general agreement about the basic features of MT evaluation are not, at the outset, subject to much dissention, but there are no collectively acknowledged and reliable methods and measures, and evaluation methodology has been the subject of much discussion in recent years.

"As in other areas of NLP [5], three types of evaluation are recognized: **Adequacy Evaluation** to determine the fitness of MT systems within a specified operational context; **Diagnostic Evaluation** to identify limitations, errors and deficiencies; and **Performance Evaluation** to assess stages of system development or different technical implementations. Adequacy evaluation is typically performed by potential users and/or purchasers of system; diagnostic evaluation is the concern mainly of researchers and developers; and performance evaluation may be undertaken by researchers, developers, or potential users.

MT evaluations typically include features not present in evaluations of other NLP systems. The quality of the raw translations, e.g., intelligibility, accuracy, appropriateness of style/register; the usability of facilities for creating and updating dictionaries, for post-editing texts, for controlling input language, for customization of documents, etc.; the extendibility to new language pairs and/or new subject domains; and cost-benefit comparisons with human translation performance."

According to Hutchins and Somers [4] the most obvious tests of the quality of a translation are:

A. **Accuracy**, that is the extent to which the translation accurately renders the meaning of the source text, without intensifying or weakening any part of the meaning [7]; and
B. **Transparency**, which is the extent to which the translation appears to a native speaker of the TL to have originally been written in that language, and conforms to the language's grammatical, syntactic and idiomatic conventions [7].

The evaluation made in this research focused on the quality of the output, i.e., the translation of two prevalent English-Persian MT programs.

[1] It is the evaluation of an MT program. There is no simple or unique way of conducting an MTE.

2. METHODOLOGY OF THE STUDY

Two prevalent English-Persian MT programs (Pars Translator & Padideh Translator) were selected as the subject of this research. The instruments of the research included a computer, a test suite[2] and detailed criteria for measuring the output.

The criteria for selecting these MT programs were:

a) Commercially available in Iran or accessible on the Internet
b) Presently and popularly used
c) Fully-automated

Given that the sentence is the basic unit in the translation process of the two programs evaluated, it was chosen as the basic testing item. Hence this research only evaluated the output quality of sentences.

Altogether there were 451 sentences, containing 282 testing points and covering 9 subjects, namely lexical coverage, phrase, morphology, simple sentences, complex sentences, syntactic ambiguity and semantic analysis, generation of Persian, special difficulties in English-Persian machine translation, and long sentences.

The test suite borrowed from the language discrete-point testing method, which means each testing item (i.e., sentence) in the suite contains a testing point. It consists of 3 parts:

- Testing an MT system' ability to analyze SL (from subject 1 to subject 6)
- Testing an MT system's ability to synthesize TL (subject 7)
- Testing an MT system's ability to deal with special difficulties in MT (subject 8 and subject 9)

The test suit I employed in the present research was originally in Chinese. The present English version was translated from Chinese by Yan Weiwei [9] served as a useful and instructive model and fitted well in with the research.

According to what is concerned with measuring accuracy, the focus was on the preservation of meaning, which involves the comparison of meaning in the output with that in the original.

Ke ping [6] distinguishes three kinds of meaning in translation in a socio semiotic approach, based on which (excluding those meanings that are impossible in or irrelevant to MT) accuracy was measured in two

[2] In Natural Language Processing (including Machine Translation), a test suite is a set of points, artificially constructed and designed to probe the system's behavior with respect to some particular language phenomena. [5]

dimensions: **referential meaning** (at lexical level, the lexical meanings of the words and phrases) and **grammatical meaning** (inflectional morphology and syntax)**.** Grammatical categories include tense, aspect, case, gender, mood, number, person, and voice.

In this research the measure of intelligibility revolved around two dimensions, i.e., grammaticality and fluency.

The scoring procedure embraced two types: The first step was to decide whether the meaning of a translation is completely unfaithful to that of the original or totally unintelligible. In either case, the translation will score zero.

In the second step, those translations not having scored zero, were assessed a set of scoring criteria as follows:

The full mark of each translation was 10 points: 5 for accuracy and 5 for intelligibility. Different weights were assigned to referential meaning, grammatical meaning, grammaticality, and fluency. Every translation lost points according to the nature and number of the errors it made. For an error at 'referential meaning at sentence level', 1.0 point was missing. Once an error was made at the level of syntax, 0.5 point was subtracted. The same weight was also given to 'word order', collocation, and idiomaticity. For those below the sentence level, i.e., lexical meaning errors, wrong tenses, etc, 0.25 was subtracted. Punctuation also lost 0.25 point. After all the points caused by all the errors were subtracted from the 10 points, the remaining points were the final score that reflected the overall quality of the output.

All the sentences in the suite were translated on computer by two MT programs. The output was gathered for further analysis. Then the criteria for measurement were applied to score the output of each program, and whether the programs had correctly translated each testing points.

The average of the total scores and the average of each subject were calculated. Then the overall rates of correction of the translation of the testing points and the rates of the correctness on each subject were also calculated. The average of the total scores and the overall rates of correctness were compared to answer the first question of the research. The rates and the means on each subject were compared and the translation of the testing points was analyzed to answer the second research question.

3. RESULTS & DISSCUSSION

The findings about the output quality of these two programs are as follows:

a) It was reflected that Padideh Translator scored a higher rate of correctness (with a slight difference of 0.5%).
b) The major linguistic bottlenecks of these MT programs in translating testing points were related to Morphology, Complex Sentences, Syntactic Ambiguity & Semantic Analysis, Generation of Persian, and Long Sentences.

3.1. Morphology

In Pars:

- Infinitives as adverbials
- The passive voice and different tenses of the Infinitive
- The perfect form and the passive voice of present participle

In Padideh:

- Special usages of comparative degree of adjectives
- The addition of *–est* to form the superlative degree of some adverbials
- The addition of preceding *most* to form superlative degree of some adverbials
- Infinitive as subject
- The perfect form and the passive voice of a gerund

Shared in Padideh and Pars:

- The addition of a preceding *more* to form the comparative degree of adverbials
- The base form, the past form, and the past participle of a verb are the same
- The meaning and translation of structure *too+ adjective/adverb+ infinitive*
- Present participle phrases as attributive placed behind the noun it modifies
- Present participle as adverbial denoting time, result, reason, condition and purpose (with the same function as that of a sentence or clause)

- Gerund as attributive
- The complex construction of gerund as subject, object, prepositional object and predicative

3.2. Complex Sentences

In Pars:

- Subject clause introduced by *what*
- Subject clause introduced by *who* or *whoever*
- Adverbial clause of reason
- Adverbial clause of manner
- Predicative clause introduced by *why*

In Padideh:

- Subject clause introduced by *when*
- Predicative clause introduced by *who* or *whom*
- Adverbial clause of comparison
- Attributive clause introduced by *preposition* + *which* or *preposition* + *whom* construction
- Appositive clause

Shared in Padideh and Pars:

- Attributive clause introduced by *which, who, whose* or *whom*
- Subject clause introduced by *where* or *wherever*
- The attributive clause is also a complex sentence.
- Neither of the two clauses in one sentence is embedded in the other clause.

3.3. Syntactic Ambiguity & Semantic Analysis

In Pars:

- Word belonging to adjective and verb
- Complex sentences which are semantically compound ones

In Padideh:

- Word belonging to adverbial and adjective
- Word belonging to adverbial and preposition
- Word belonging to noun, adjective and verb
- Word belonging to conjunction, adjective and verb
- Word belonging to demonstrative pronoun, relative pronoun and subordinate conjunction
- Nouns with different meanings
- Subordinate conjunctions which have various meanings and introduce different types of clauses

Shared in Padideh and Pars:

- Word belonging to pronoun and relative pronoun
- Word belonging to conjunction and preposition
- Word belonging to conjunction and adverb
- Verbs with different meanings in accordance with what follows
- To judge whether the present participle helps to form the predicate verb or act as a nominal modifiers

3.4. Generation of Persian

In Pars:

- The definite articles are often omitted.
- The translation of negative imperative sentences
- Not only the word order of post-attributives of nouns and pronouns but also the word order within these modifiers should be adjusted
- The word order of the translations should be adjusted when some attributives are placed behind the nouns or pronouns it modifies
- Logical indirect object+ passive form of verb+ by+ logical subject

In Padideh:

- Negative sentences should be translated into affirmative ones
- Logical direct object+ passive form of verb+ to+ indirect object+ by+ logical subject

Shared in Padideh and Pars:

- The indefinite article a before a noun as a unit of measure should be translated into "یک"/yek/ (i.e., one) or "هر" /har/ (i.e., any)

3.5. Long Sentences

In Pars:

- Complex sentences with layer of subordination

Shared in Padideh and Pars:

- Simple sentences with many or long modifiers

4. CONCLUSION

The researcher arrived at the following conclusions concerning the output quality of Two English_Pesian MT programs:

A. Padideh Translator produces the best output.
B. Based on the most unsuccessfully translated testing points, the major linguistic bottlenecks of these MT programs were in the areas of Morphology, Complex Sentences, Syntactic Ambiguity & Semantic Analysis, Generation of Persian, and Long Sentences. These 5 subjects are mainly connected with the complexity in the syntactic and semantic analysis, or the difference between English and Persian.

Both programs performed fairly well in their treatment of relatively simple language phenomena, i.e., lexical coverage, phrases, simple sentences, and, to a lesser extent, morphology, but it was not the same case with their performance on the relatively complicated language phenomena, especially syntactic ambiguity and long sentences, and the generation of the target language as well. The main reason may possibly lie in the imperfection of their translation engine, which failed to take many complicated language phenomena and other difficulties in MT into consideration. In fact, the designers should have mentioned the following recommendations (or disclaimers) in the user's guides:

- Use short, declarative sentences. Declarative sentences consist of subject, verb, and object, in that order. Imperatives, wordy or convoluted sentences, and some types of questions are difficult to analyze. Sentences composed of phrases linked by conjunctions may also produce mistranslations.
- Use unambiguous words. The dictionaries allow only one translation for a word or phrase, avoid using words whose most common meaning is not the one you intended; instead, find a synonym for the specific meaning you want. For example the word *head*, when used as a noun, has several meanings. The most common meaning is the part of the body that contains the brain; director or leader is another meaning. If you mean director use the word *director* rather than *head*.
- Avoid idiomatic or informal expressions, unless you add them to the Semantic Unit Dictionaries.
- Include redundant relative pronouns, prepositions, and other words that clarify the sentence structure.

5. ACKNOWLEDGMENTS

I especially would like to offer my sincerest appreciation for great contribution of Mrs. Maria Teresa Costa Gomes Roberto, the Auxiliary Professor of the University of Aveiro, in preparing this article.

6. REFRENCES

[1] Almasoud, 'A Machine Translation (MT) or Mad Translation?' Presentation at the International Conference of Translators and Computers Session, ATA November 4-8, 1998.

[2] Craciunescu, et al. (2004). Machine Translation and Computer-Assisted Translation: a New Way of Translating? Translation Journal (TJ), Vol. 8, No. 3

[3] Examples of Nice Translation Made by Automated Translators Available on the Market' http://www.fortunecity.com/business/reception/19/mtex.htm

[4] Hutchins, John and Somers, Harold (1992). *An Introduction to Machine Translation*. London: Academic Press Limited.

[5] Hutchins, J. (1997). Evaluation of Machine Translation and Translation Tools. http://www.hutchinsweb.me.uk/HLT-1997.pdf

[6] Jurafsky, D. & Martin, J. H. Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition. University of Colorado, Boulder: Prentice Hall. 2000.

[7] Ke Ping. A socio semiotic Approach to Translation. Vol. 9 N0. 3, February 2006.

[8] Kulesza, A. & Shieber, S. M. A learning Approach to Improving Sentence Level MT Evaluation, in Proceedings of the Tenth Conference on Theoretical and Methodological Issues in Machine Translation, Baltimore, 2004.

[9] Weiwei, Y. Evaluation of the Output Quality of some Prevalent English-Chinese Machine Translation Programs. http://freewebs.com/keping/P-MA-2004-YWWEVOfTheOutputQualityOfSomePrevale