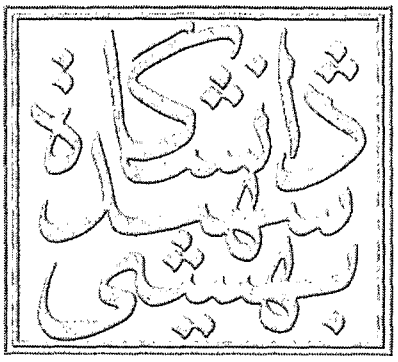


Shahid Beheshti University

Faculty of Letters and Humanities

Department of English

The Effect of Pre-editing on Output Quality of Machine Translation

A Thesis Submitted in Partial Fulfillment of the Requirements for the Master of Arts in English Translation (Translation Studies)

By: Mojtaba Hatef

Thesis Supervisor: Dr. S. Baleghizadeh

Thesis Reader: Dr. A. Fatemi

Tehran, Iran

September 2009

فرم اطلاعات پایان نامه های کارشناسی ارشد

عنوان پایان نامه: تاثیر پیش ویرایش در کیفیت خروجی ترجمه ماشینی	
نام دانشجو: مجتبی	**نام خانوادگی دانشجو:** هاتف
رشته تحصیلی: مترجمی زبان انگلیسی	**شماره دانشجویی:** ۸۵۴۱۱۱۸۴
استادراهنما: دکترساسان بالغی زاده	**استاد مشاور:** دکتر سیدابوالقاسم فاطمی جهرمی
استاد داور: دکتر محمدرضا عنانی سراب	**تعداد واحد پایان نامه:**۴
تاریخ شروع پایان نامه:	**تاریخ اتمام پایان نامه:**
تاریخ دقیق دفاع از پایان نامه:	**نمره پایان نامه به عدد:** ۱۹
نمره پایان نامه به حروف: نوزده	**معادل:** عالی
آدرس و شماره تلفن دانشجو: استان اردبیل-شهرستان پارس آباد مغان-دهستان اسلام آباد قدیم-تلفن ۰۴۵۲۷۴۶۳۲۱۳	

چکیده پایان نامه:

اهداف:

بررسی تاثیر پیش ویرایش در کیفیت خروجی ماشین ترجمه پدیده

روش های اجرا:

استفاده از مدل ارزیابی کارول برای ارزیابی کیفیت خروجی ماشین ترجمه پدیده قبل و بعد از پیش ویرایش

خلاصه پایان نامه: ابتدا متن های خروجی نرم افزار پدیده جهت مشخص کردن مشکلات متداول زبانی این نرم افزار مورد بررسی قرار گرفت، سپس با توجه به مشکلات عمده، پیش ویرایش مناسب تعریف گردید. پس از اعمال پیش ویرایش، کیفیت خروجی ماشین ترجمه پدیده قبل و بعد از پیش ویرایش از لحاظ قابلیت فهم و صحت مورد ارزیابی قرار گرفت.

نتایج:

نتایج تحقیق نشان داد که پیش ویرایش، بهبود قابل ملاحظه ای در کیفیت خروجی ماشین ترجمه پدیده بدست نمی دهد.

کلید واژه ها:

۱.هوش مصنوعی ۲.صحت ۳.قابلیت فهم ۴.ترجمه ماشینی ۵.پردازش زبان طبیعی ۶. پیش ویرایش

استاد راهنما:	**استاد مشاور:**	**استاد داور:**
نام و نام خانوادگی: ساسان بالغی زاده	**نام و نام خانوادگی:** سید ابراهیم فاطمی	**نام و نام خانوادگی:** محمدرضا عنانی سراب
امضاء:	**امضاء:**	**امضاء:**

۱۳۸۸/۱۰/۲۷

تاریخ اطلاعات مدارک علمی دانشگاه

تهیه مدارک

۱۲۹۴۲۵

Abstract

Machine Translation as an inevitable and increasing demand of the day needs to be utilized and adapted to meet the needs of our daily-modernized lives. Regarding this, the present research is an attempt to study the effect of pre-editing on output performance of an English-Persian machine translation system, Padideh Translator. To conduct the research two major stages were set: at the first stage, the researcher explored the output texts to decide on the system's major and recurrent linguistic errors in order to define some pre-editing strategies to lighten their effects, and secondly, he investigated the effectiveness of pre-editing in producing higher quality output by the system. Then, the researcher conducted a declarative testing to evaluate the effectiveness of Padideh machine translation (MT) system. Output quality criteria included intelligibility and fidelity that were measured and compared using pointed scales. The research showed that pre-editing does not improve significantly the output quality of Padideh Translator in terms of intelligibility and fidelity.

Key words: artificial intelligence, fidelity, intelligibility, machine translation, natural language processing, pre-editing.

Acknowledgements

Upon finishing this thesis, I felt indebted to a group of people, without whom this work would have never been completed. I would express my sincere gratitude to all of them. I feel especially obliged to give my warm thanks to Dr. Baleghizadeh, my thesis supervisor, and Dr. Fatemi, my thesis reader, without whose invaluable contribution I would never have had the opportunity to complete this work.

Table of Contents

TITLE PAGE 1

ABSTRACT 2

ACKNOWLEDGEMENTS 3

TABLE OF CONTENTS 4

LIST OF ABBREVIATIONS 6

CHAPTER ONE: INTRODUCTION 7

1.1 STATEMENT OF THE PROBLEM 8
1.2 SIGNIFICANCE OF THE STUDY 9
1.3 PURPOSE OF THE STUDY 10
1.4 RESEARCH QUESTIONS 11
1.5 RESEARCH HYPOTHESIS 11
1.6 THEORETICAL FRAMEWORK 11
1.6.1 Fully Automatic High Quality Translation (FAHQT) *13*
1.6.2 Machine-Aided Human Translation (MAHT) *13*
1.6.3 Human-Aided Machine Translation (HAMT) *15*
1.6.3.1 Pre-editing 15
1.6.3.2 Post-editing 16
1.6.3.3 Interactive MT 17
1.7 DEFINITION OF KEY TERMS 18
1.7.1 Artificial Intelligence (AI) *18*
1.7.2 Computational Linguistics *18*
1.7.3 Controlled Language *18*
1.7.4 Machine-aided (Computer-aided) Translation (MAT) *18*
1.7.5 Machine Translation (MT) *19*
1.7.6 Machine Translation Evaluation (MTE) *19*
1.7.7 MT Output *19*
1.7.8 Natural Language Processing (NLP) *19*
1.7.9 Pre-editing *20*
1.8 SCOPE AND LIMITATIONS OF THE STUDY 20

CHAPTER TWO: LITERATURE REVIEW **22**

2.1 SIGNIFICANCE OF MT 23
2.2 A BRIEF HISTORY OF MT 26
2.3 MT PARADIGMS 35
2.3.1 Rule-based *35*
2.3.1.1 Direct Translation 36
2.3.1.2 Indirect Approach 37
2.3.2 Corpus-based MT *39*
2.3.2.1 Statistics-based MT 40
2.3.2.2 Example-based MT 40
2.3.3 Hybrid Systems *41*
2.4 STATE-OF-THE-ART MT SYSTEMS 42
2.5 MACHINE TRANSLATION EVALUATION (MTE) 44

2.5.1 General Types of MT Evaluation *47*
2.5.1.1 Feasibility Testing 48
2.5.1.2 Requirements Elicitation 48
2.5.1.3 Internal Evaluation 49
2.5.1.4 Diagnostic Evaluation 49
2.5.1.5 Declarative Evaluation 50
2.5.1.6 Operational Evaluation 51
2.5.1.7 Usability Evaluation 51
2.5.2 MT Evaluation Methods *52*
2.5.2.1 Carroll's Model (1966) 52
2.5.2.2 Crook & Bishop's Model (1979) 54
2.5.2.3 Sinaiko's Model (1979) 54
2.5.2.4 Nagao's Model (1985) 55
2.5.2.5 DARPA Model (1992-1994) 56
CHAPTER THREE: METHODOLOGY **59**
3.1 RESEARCH DESCRIPTION 60
3.2 MATERIALS AND INSTRUMENTS 60
3.2.1 Padideh Translator *60*
3.2.2 Microsoft Office Word 2007 and Microsoft Office Excel 2007 ..... *61*
3.2.3 Scoring Model *62*
3.2.3.1 Measurement of Intelligibility 62
3.2.3.2 Measurement of Fidelity 64
3.2.4 Sample Texts *67*
3.3 PROCEDURES 67
3.4 DATA COLLECTION 69
3.5 DATA ANALYSIS 69
CHAPTER FOUR: RESULTS AND DISCUSSION **70**
4.1 FINDINGS 71
4.1.1 Some Major Problems of Padideh Translator & Pre-editing Strategies to Lighten their Effects *71*
4.1.2 Results of Scoring *87*
4.2 DISCUSSION 95
CHAPTER FIVE: CONCLUSION **99**
5.1 SUMMARY 100
5.2 IMPLICATIONS OF THE RESEARCH 102
5.3 SUGGESTIONS FOR FURTHER RESEARCH 103
REFERENCES **105**
APPENDIX A **110**
APPENDIX B **145**

List of Abbreviations

AI: Artificial Intelligence

CAT: Computer-Aided (or Computer-Assisted) Translation

FAHQMT: Fully Automatic High Quality Machine Translation

FAHQT: Fully Automatic High Quality Translation

HAMT: Human-Aided Machine Translation

MAHT: Machine-Aided Human Translation

MT: Machine Translation

MTE: Machine Translation Evaluation

NLP: Natural Language Processing

CHAPTER ONE: INTRODUCTION

Not very long ago, automatic translation between human natural languages had been a matter of science fiction and a scientific dream. However, today, owing to the rapid growth of Computational Linguistics and Artificial Intelligence, this dream has come true; though not to the degree that had been seen in science-fiction world. Although there still remain major problems in this field, "some degree of automatic translation is now a daily reality" (Arnold, Lorna, Siety, Humphreys, & Sadler, 1994, p. i). According to Arnold et al. (1994), machine translation (MT) is now an "important topic — socially, politically, commercially, scientifically, and intellectually or philosophically" (p. 4). In the last few decades, research on MT has made major progress. One of the subfields of this area that is essential for its progress is Machine Translation Evaluation (MTE), which tries to evaluate MT approaches and systems regarding various aspects related to this area. Concerning these, this research is a study about the effectiveness of human and machine interaction in the field of MT.

1.1 Statement of the Problem

Since the early days of MT life, there have been big claims and expectations about it. On the other hand, there are those who, assessing MT using the same criteria of human translation assessment and referring to failures of MT in the field of literary translation, believe that MT is absolutely useless and there may be no prospect for it. However, the growing demands for technical translations and shortage and high costs of human translators with low speed have made the use of MT in this technological and scientific era inevitable. According to Wilks (2009), only in Japan, many thousands of MT systems have been sold, and every day millions of people around the world click the [Translate this page] option on web browsers and see

immediate gain and benefit from what it gives back. In addition, he argued that "the absence of any intellectual breakthroughs to produce indisputably high-quality fully-automatic MT is equally clear, a fact which has led some to say it is impossible, a claim inconsistent with our first observations" (Wilks, 2009, p.1). Today, we can see domain specific MT programs, such as machine translation closed captions (Popowich, McFetridge, Turcato, & Toole, 2000). Therefore, it is clear that today MT is a necessity rather than a possibility but what is to be done in this domain is to make it as useful as possible for our daily requirements. As Hutchins & Somers (1983) stated:

> What counts as a 'good' translation, whether produced by human or machine, is an extremely difficult concept to define precisely. [...] What matters in practice, as far as MT is concerned, is how much has to be changed in order to bring output up to a standard acceptable to a human translator or reader. (p. 2)

Accordingly, they conclude that with a slippery concept such as translation, researchers of MT systems can finally aspire only to producing translations which are 'useful' in particular situations or, alternatively, seek suitable applications of the 'translations' which in fact they are able to produce. Consequently, MT as an inevitable and increasing demand of the day needs to be adapted and utilized to meet the needs of our daily-modernized lives.

1.2 Significance of the Study

Today, MT may be considered as a necessity rather than a possibility but according to Hutchins & Somers (1983) "what matters in practice, as far as MT is concerned, is how much has to be changed in order to bring output up to a standard acceptable to a human translator or

reader" (p. 2). MT systems may be ranked according to their degree of automaticity. A system may be fully automatic or interactive; if interactive, it may provide basically human translation with some assistance from the machine, or basically machine translation with some assistance from the human translator. When the fully automatic MT is not appropriate, "the performance of the MT system can be mitigated by the use of pre- or post-editing, i.e. adapting the input or output text to meet the end-user's needs" (Somers, 1998a, p. 138). Thus, in evaluating MT systems, it may be helpful to determine the degree of pre-editing effectiveness and, consequently, the degree of human and machine interaction. Regarding these considerations, this thesis is an attempt to study the effect of pre-editing on output performance of an English-Persian MT system, Padideh Translator.

1.3 Purpose of the Study

The aim of this research is to investigate the effect of pre-editing on output quality of an English-Persian MT system, Padideh MT system which is commercially known as Padideh Translator. The purpose of this study is, firstly, to determine the linguistic deficiencies in the output of the evaluated MT system in order to define some pre-editing strategies to lighten the bad effects of the problems, and secondly, to investigate the effect of defined pre-editing strategies (regarding the deficiencies) to improve the mentioned system's output quality. This study intends to show that whether pre-editing improves the MT system's output quality or not.

1.4 Research Questions

1. What are the main linguistic deficiencies of Padideh MT system that make the output quality of the system poor?
2. What strategies can be taken as pre-editing techniques to improve the output quality of the MT system?
3. Do the pre-editing strategies improve the output quality of the system?

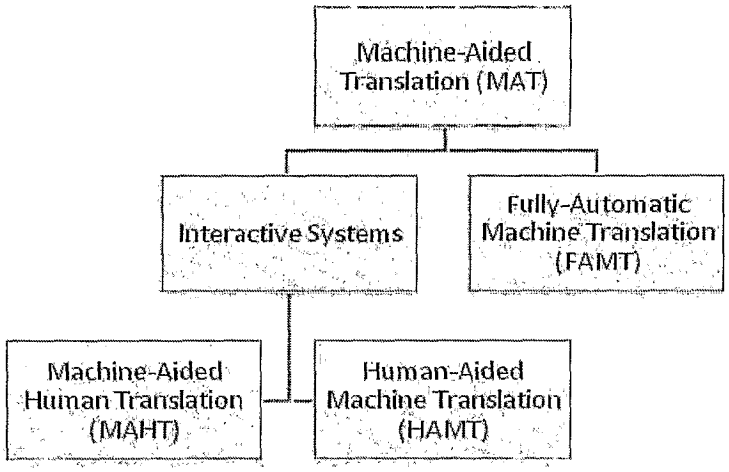
1.5 Research Hypothesis

- Pre-editing strategies improve the output quality of Padideh MT system.

1.6 Theoretical Framework

MT systems may be ranked according to their degree of automaticity. A system can be fully automatic or interactive; if interactive, it may provide basically human translation with some assistance from the machine, or basically machine translation with some assistance from the human translator. According to Lehrberger & Bourbeau (1998, p. 201), this yields the following classification (Figure 1-1):

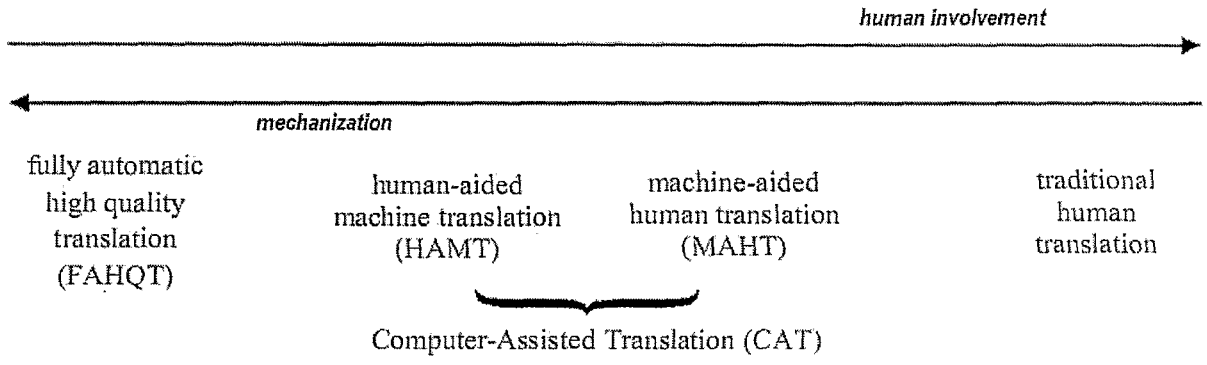
Figure 1-1 Lehrberger & Bourbeau classification of degree of translation automacity (Lehrberger & Bourbeau, 1998, p. 201)



There is no clear-cut distinction between the boundaries of Machine-Aided Human Translation (MAHT) and Human-Aided Machine Translation (HAMT) and the term Computer-Aided (or Computer-Assisted) Translation (both CAT) is sometimes used to refer to both; an umbrella term covering a continuum from human translation in the proper sense of the word to fully automatic machine translation (Freigang, 1998). However, the central core of MT itself is the automation of the full translation process (Hutchins & Somers, 1983).

Examining the modes of use of MT, Hutchins & Somers (1983) suggested a similar spectrum (represented in Figure 1-2) showing various kinds of human-machine cooperation. At one extreme are wholly computerized systems with no human involvement producing translations of a high quality: namely fully automatic high quality translation (FAHQT). At the other extreme is human translation involving no computerized aids whatever as it has been practiced for centuries. Between them are Human-Aided Machine Translation (HAMT) and Machine-Aided Human Translation (MAHT).

Figure 1-2 Human & machine translation (Hutchins & Somers, 1983, p. 148)



1.6.1 Fully Automatic High Quality Translation (FAHQT)

The term "fully automatic high quality translation" (FAHQT) was originally coined by Yehoshua Bar-Hillel (1951, as cited in Hutchins & Somers, 1983). He always argued that fully automatic translation of a quality comparable to that of human translations was an unrealistic aim and impossible in principle. His 1960 report as "the present status of automatic translation of languages" was an attempt to direct MT to the more realistic aims of "Human-Aided Machine Translation" (Bar-Hillel, 1960/2003).

1.6.2 Machine-Aided Human Translation (MAHT)

As it is evident from the classification in Figure 1-2, Machine-aided translation (also known as computer-aided translation) may be divided into two different subcategories: Machine-Aided Human Translation (MAHT) and Human-Aided Machine Translation (HAMT). MAHT involves utilization of machine aids in the process of translation by human

translators. MAHT might be considered as a system in which the machine is used as a tool at the service of the human translator. According to Blatt et al. (1985, as cited in Freigang, 1998), machine aids can cover systems such as word processors, dictionary management tools, term banks, and various look-up facilities which support the translator but do not actually perform the translation task. In Freigang's view (1998), MAHT covers standard software systems used in a modern office environment in general rather; these include standard word-processing software, universal database systems and other tools used in performing administrative tasks. The systems discussed under this category "are not designed to undertake any syntactic or semantic analysis of a source text nor to generate a target language equivalent of the source text or any part of it" (Freigang, 1998, p.134). Lehrberger & Bourbeau (1998) identified six features that may be included in an MAHT system:

(i) Word processor with provision for dictionary lookup (translation equivalents).

(ii) KWIC facility. The KWIC (Key Word In Context) can be used to show the contexts in which a word occurs in the texts under translation or in texts from the same domain. This helps the translator to understand how a word is used in that domain and may therefore help in the [ambiguity] resolution of homographs.

(iii) Grammatical information. In addition to providing translation equivalents, the machine might also supply, for each word in its dictionary, grammatical categories (i.e., parts of speech), sub-categories, and various syntactic and semantic properties of the word. [...]

(iv) Morphological analysis.

(v) Corpus of translated texts. The translator can be provided with easy access to previously translated texts for reference in the current task.

(vi) Spelling and grammar correction. (p. 6)

1.6.3 Human-Aided Machine Translation (HAMT)

While in MAHT it is the human who is in charge of translation proper, in HAMT the system itself takes the main responsibility for translation, with some human assistance in the overall process of translation. This assistance can be accomplished at several stages; it may be before machine processing begins (pre-editing), during the process (interactive MT), or after the process (post-editing).

1.6.3.1 Pre-editing

Often, a human translator can turn a badly written text into a well written translation; but this is not true about an MT system. According to Kumar (1994), at pre-editing stage, "the editor (operator) intervenes before translation to eliminate lexical and structural ambiguities by either revising the text by editing software, or by customizing the text for translation according to pre-established rules and vocabulary" (p. 6). This may not eliminate the post-editing phase but can make it easier for the system and may improve the output in terms of time-effectiveness of post-editing. Pre-editing can include the identification of proper nouns, the marking of grammatical categories of homographs, indication of embedded clauses, bracketing of coordinate structures, and substitution of unknown words (Hutchins & Somers, 1983). Pre-editing, in its advanced form, involves revision of texts using restricted input, that in its extreme form is called controlled language. For example, Arnold et al. (1994) have

suggested some simple writing rules and strategies that can improve the performance MT systems:

- Keep sentences short.
- Make sure sentences are grammatical.
- Avoid complicated grammatical constructions.
- Avoid (so far as possible) words which have several meanings.
- In technical documents, only use technical words and terms which are well established, well defined and known to the system. (p. 26)

The restricted input scenario might be very advantageous in the case of machine translation of domain-specific texts into multiple languages (Somers, 1998a).

1.6.3.2 Post-editing

Post-editing is currently the most common human intervention in MT process. It "consists of tidying up the raw output, correcting mistakes, revising entire, or, in the worst case, retranslating entire sections" (Somers, 1998a, p. 138). However, post-editing has its own negative aspects; for example, as Somers (1998a) pointed out, correcting MT output was quite different from revising human output, and many translators found it frustrating. In addition, since the initial quality is quite low, repair can take longer and may be more difficult than simply starting from scratch. Due to frequent mistakes made by MT systems, some advanced systems provide their users with specific interactive tools especially designed to make post-editing easy. For example, Padideh Translator comes with an interactive mode that makes it possible for post-editor (usually the translator) to edit the output text sentence by sentence.

Moreover, with this interactive mode, the system suggests some alternatives for ambiguity resolution. A similar feature is available in Pars Translator, though named Pars Editor.

Although most post-editors are translators and are accustomed to producing high quality texts, "some MT output could be subject to a rough and ready post-edit — where the post-editor tries to remove or adjust only the grossest errors and incomprehensibilities — rather than the usual thorough and painstaking job" (Arnold et al., 1994, p.32). The major advantage of this option, for example in the case of emails, is its time effectiveness.

1.6.3.3 Interactive MT

Interactive MT proper is distinguished from interactive pre- or post-editing. Whereas pre-editing and post-editing is done interactively, as Hutchins & Somers (1983) cited, interactive MT "refers strictly to human involvement during the actual process of translation (analysis, transfer, and generation) when the computer seeks assistance in the interpretation of structures, the resolution of ambiguities and the selection of lexical items" (p. 151). According to Somers (1998a), interactive approach is now losing popularity because of some drawbacks. The main drawbacks are that the user must have some knowledge of both languages and since MT systems typically translate sentence-by-sentence, the system would ask questions as it comes to them, rather than arranging for related questions to be asked all together. In addition, as the interactions are usually "canned" texts, the questions tend to be very repetitive, and the system may ask exactly the same question several times during the course of translation. Considering these as well as the fact that the user in typically a translator, interactive systems have been found to be too much slower than manual translation.

1.7 Definition of Key Terms

1.7.1 Artificial Intelligence (AI)

Artificial Intelligence (AI) is referred to the branch of Computing Science concerned with simulating aspects of human intelligence such as language comprehension and production, vision, planning, etc (Arnold et al., 1994).

1.7.2 Computational Linguistics

"Computational linguistics is a term applied to any type of computer-assisted treatment of natural languages" (Lehrberger & Bourbeau, 1998, p. 2).

1.7.3 Controlled Language

Controlled language is:

> A specially simplified version of a language which is adopted (typically by a company or a documentation section of a company) as a partial solution to a perceived communication problem. Both the vocabulary and the syntactic structures may be restricted. (Arnold et al., 1994, p. 201)

1.7.4 Machine-aided (Computer-aided) Translation (MAT)

Machine-aided translation (also known as computer-aided translation) is an umbrella term covering a continuum from human translation in the proper sense of the word to fully automatic machine translation (Freigang, 1998).

1.7.5 Machine Translation (MT)

According to Arnold et al. (1994), Machine Translation (MT), or sometimes called Automatic Translation, is a sub-filed of computational linguistics that deals with the use of computer in translating natural languages, to automate all, or part of the process of translating from one human language to another.

1.7.6 Machine Translation Evaluation (MTE)

Machine translation evaluation (MTE) is a field of MT that assesses various aspects of MT systems including linguistic evaluation of raw output, technical evaluation by researchers and developers and evaluation by potential users (Hutchins & Somers, 1983).

1.7.7 MT Output

MT output is referred to the translation (target language text) of the input (source language text) produced by MT system.

1.7.8 Natural Language Processing (NLP)

Natural Language Processing (NLP) is referred to the field of inquiry concerned with the study and development of computer systems for processing natural (human) languages (Arnold et al., 1994).