Qualitative Analysis of Cotemporary Urdu Machine Translation Systems

Asad Abdul Malik, Asad Habib

Kohat University of Science and Technology, Kohat, Pakistan

asad_12204@yahoo.com, asadhabib@kust.edu.pk

Abstract. The diversity in source and target languages coupled with source language ambiguity makes Machine Translation (MT) an exceptionally hard problem. The highly information intensive corpus based MT leads the MT research field today, with Example Based MT and Statistical MT representing two dissimilar frameworks in the data-driven paradigm. Example Based MT is another approach that involves matching of examples from large amount of training data followed by adaptation and re-combination.

Urdu MT is still in its infancy due to nominal availability of required data and computational resources. This paper provides a detailed survey of the aforementioned contemporary MT techniques and reports findings based on qualitative analysis with some quantitative BLEU metric quantitative results. Strengths and weaknesses of each technique have been brought to surface through special focus and discussion on examples from Urdu language. The paper concludes with proposal of future directions for research in Urdu machine translation.

Keywords: Urdu Machine Translation, Qualitative Comparison, Rule Based MT, Statistical MT, Example Based MT

1 Introduction

Representing text in one natural language, the source language (SL) into another, the target language (TL) is as old as the written literature [1]. At present, the need of translation is continuously growing in business, economy, medical and many other fields. The growth in science and technology in general and computer based solutions in particular have paved the way to the concept of automatic translation called the Machine Translation (MT) [2].

1.1 Urdu

Urdu ranks 19th among the 7,105 languages spoken in the world[1]. It is one the most-spoken languages in South Asia [3]. It is also spreading in the West due to the large

[1] http://www.ethnologue.com/statistics/size

Diaspora of Indo-Pak Subcontinent citizens. Urdu is the national language of Pakistan and it is used i) as medium of teaching in most of the public schools ii) for junior to mid level administration and iii) in the mass print and electronic media. It is not only spoken in Pakistan but also in India, Bangladesh, Afghanistan and Nepal. Also it has become the culture language and lingua franca of the South Asian Muslim Diaspora outside the Indo-Pak subcontinent, mainly in the Middle East, Europe, Canada and the United States [4].

1.2 Urdu Machine Translation (UMT)

In spite of the large number of speakers around the world, there are very few computational natural language tools available for Urdu. It is a morphologically rich language having many other distinct linguistic characteristics. On the contrary it is still an under-resourced language from the point of view of computational research. We could not find any public domain machine translation tool(s) developed specifically for Urdu. However some trace of basic MT techniques has been discovered [5-9]. In the current work we presented a detailed survey on the contemporary research in UMT. We identified the weaknesses and strengths of each technique and proposed the guidelines for future directions in UMT research.

2 Literature Survey

Some traces of basic UMT research are presented in this section. Naila et al [5] presented a Rule Based English to Urdu Machine Translation (RBMT) technique primarily based on the transfer approach that tries to handle the case phrases and verb postpositions using Paninian grammar. Statistical Machine Translation (SMT) between languages with word order differences was discussed by Bushra et al [6]. Example Based Machine Translation (EBMT) approach was introduced by Maryam and Asif that translates text form English to Urdu that supports idioms and homographs [7]. Parallel corpus for statistical machine translation for English to Urdu text was presented by Aasim et al [8]. Word-Order Issues in English-to-Urdu have been investigated by Bushra and Zeman [9]. In addition, SMT systems such as Google[2] and Bing[3] are already available online. However these systems offer poor translation quality and limited accuracy due to issues related to Urdu syntax and other intrinsic linguistic features.

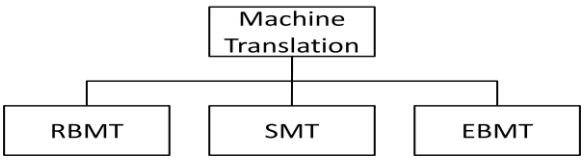

Fig. 1. Paradigms for Machine Translation

[2] http:// translate.google.com
[3] http://www.bing.com/translator

Contemporary Machine Translation techniques can be broadly categorized into three paradigms as shown in Figure 1.

2.1 Rule Based Machine Translation (RBMT)

To provide suitable rules for translation, the RBMT needs linguistic knowledge of source as well as the target language. Translation depends on formalized linguistic knowledge represented in lexicon along with grammars [10]. RBMT is described by several characteristics; it has firm set of well fashioned rules, several rules rely on present linguistic theories and the grammatical errors are prohibited. The major advantage of RBMT is that if the required knowledge is not found in available literature then ad-hoc heuristic rules are applied [5]. This system contains input sentence analyzer (morphological, syntactic and semantic analysis) and procedures for producing output (structural transfers and inherent Inter-lingua structures).

2.2 Statistical Machine Translation (SMT)

Two models are built in SMT; i) Translation model and ii) Language model. A translation model gives probability of a target sentence given source sentence P(T/S) whereas the language model determines the probability P(S) of the string of target language actually occurring in that language. By using the language model and conditional probabilities of translation model, P(S/T) is calculated using the following formula:

$$P\left(\frac{S}{T}\right) = \frac{P(S)P\left(\frac{T}{S}\right)}{P(T)}$$

Probability based analysis of MT is part of SMT. It has numerous diverse applications such as those in word sense disambiguation or structural disambiguation etc. [11]. The SMT techniques do not need explicit encoding of the linguistic information. It highly depends upon availability of fine and very large amount of bilingual data that presently does not exist for Urdu and other languages spoken in the Indo-Pak Subcontinent region.

2.3 Example Based Machine Translation (EBMT)

Somers referred to EBMT as a hybrid approach of RBMT and SMT [12]. Like SMT, it is depended upon a corpus of available translations. That is why it is similar to (often confused with) translator's aid known as Translation Memory (TM). EBMT and TM both involve comparison of input text with the database of real examples and then find out the nearest match. In TM, a translator selects the candidate target text whereas EBMT makes use of automated procedures that identify the translation fragments. Recombination of these fragments produces the target text [10].

Thus the process is split into three phases [10]. i) "Matching" fragments against the available database of real examples (that are common between EBMT and TM), ii)

"Alignment" identifying corresponding translation fragments and finally iii) "Recombination" that gives the target text. EBMT needs a database of parallel translations that are searched for source language phrases or sentences and their nearest matching target language components are generated as output [11].

EBMT saves the translation examples in different manners. In simple case, examples are saved as pairs of strings with no extra information related to them.

3 Methodology

In this section we discuss the methodologies of three major Machine Translation techniques. English is considered as source language and Urdu as the target language. We compare the strengths and weaknesses of these techniques in Section 4.

3.1 Rule Based Machine Translation

There are three stages in RBMT; i) Analysis, ii) Transfer and iii) Synthesis

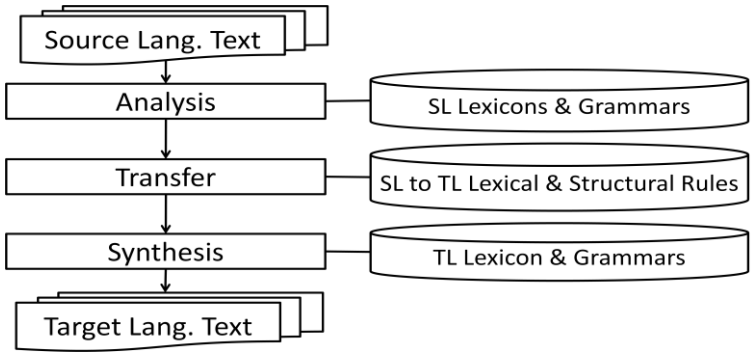

Fig. 2. RBMT Model

Analysis.
The source text is analyzed based upon lexicon and grammar rules of source language. Word segmentation is done and each word is annotated by appropriate POS tag and parse tree of input text is created. A parse tree for the input text "I called you several times" is created as shown in figure 3.

Transfer.
In this stage, parse tree of source language text is 'transferred' into parse tree of desired target language according to the lexicon and structural rules of the target language. English is SVO (Subject, Verb, Object) language whereas Urdu is SOV language. Re-ordering of words is inevitable in order to generate the output parse tree as shown in Figure 4.

English to Urdu Translation Rules.

Some coarse grained rules for translation from English to Urdu are mentioned in the following.

1. NP in both languages follows the same rule. So swapping in not required.
2. If NP is having NP and PP, then transform it as in Urdu PP comes before NP.
   English $NP \rightarrow NP + PP$
   Urdu $NP \rightarrow PP + NP$
3. If adverb phrase (AP) appears before verb then swapping is not needed. AP in English can appear in different order depending on the type of AP, however Urdu prefers AP before verb.
   Urdu AP +V
4. In Urdu, Verb phrase (VP) is inflected according to gender, number and person (GNP) of the head noun while NP depends upon tense, aspect and modality of the verb phase (VP). Urdu adjectives are also modified by GNP of the head noun.

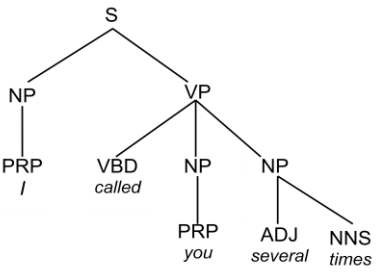

Fig. 3. English Parse Tree

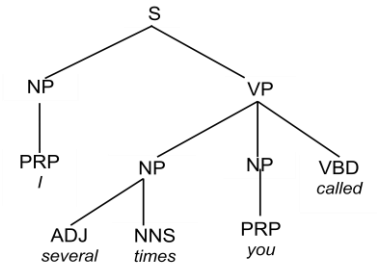

Fig. 4. Parse Tree (transferred in SOV)

Synthesis.

Finally, the target language lexicon and grammar is used to convert the parse tree of target language to the target language surface form. It requires two independent monolingual dictionaries so that appropriate surface form of target language can be generated.

As shown in figure 5 the source text “I called you several times” is translated into “میں کئی مرتبہ آپ کو بلایا” using RBMT.

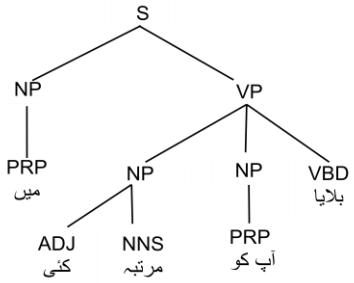

Fig. 5. Urdu parse tree

3.2 Statistical Machine Translation (SMT)

SMT makes use of i) Translation Model, ii) Language Model and iii) Decoder Algorithm.

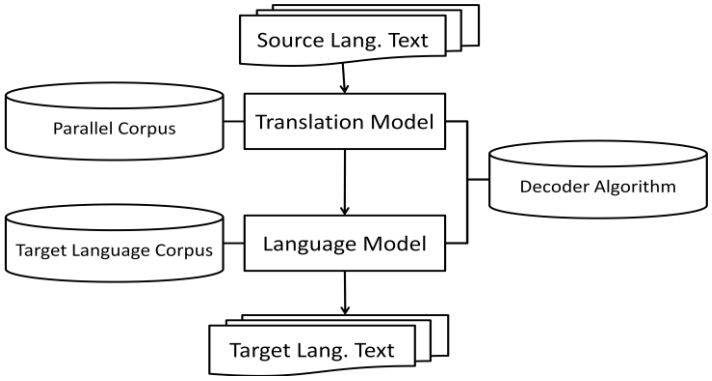

Fig. 6. SMT Model

Translation Model.
Words and phrases in the source text are matched against the target language strings. If the strings are matched the model assigns a probability value P(T/S) to it. This probability shows that what are the chances that the input text string is present in the output or target language. These probability values are pre-assigned in a parallel corpus through human translation. Subsequently machine learning techniques are used to improve the system depending upon the human translated text.

Language Model.
Language model determines the probability P(S) of output text string. It does not require a parallel corpus. It requires text in only one language. We can calculate the value by using N-gram model. In this the probability of occurrence of sentence of length N is the product of probability of each k^{th} word given the occurrence of previous words k-1 and k-2.

Decoder Algorithm.
After finding the product of translation and language model the decoder algorithm selects the string of output text language with the highest probability value based on the stochastic formula mentioned in Section 2.2.

3.3 Example Based Machine Translation (EBMT)

English to Urdu EBMT is divided into four phases; i) Sentence Fragmentation, ii) Search in Corpus, iii) N-ary Product based Retrieval and iv) Ordering of Translated Text.

Sentence Fragmentation.
For better handling of input sentence by translator, it is better to break the sentence into phrases. On the other hand same results are achieved by storing sentence in the corpus and by gaining a broad coverage by fragmenting and combining using genetic algorithm at run time for obtain new sentences. Fragmentation of a sentence into phrases is handled by using concept of idioms, cutter points and connecting words.

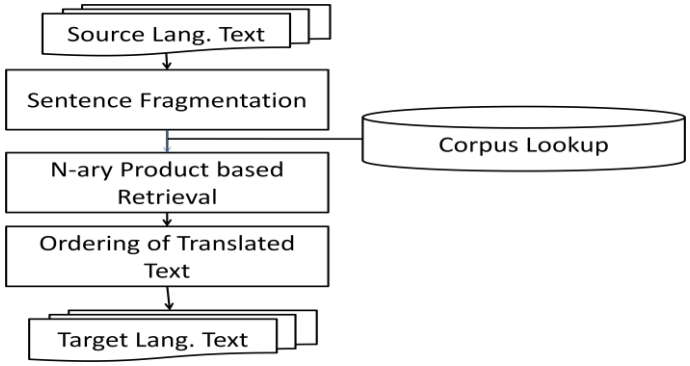

Fig. 7. EBMT Model

Searching in Corpus.
Bilingual corpus is searched for finding whether the input phrase is accessible or not. If the system is unable to locate exact match, then in that situation it will look for the nearest match. Closeness is calculated by threshold at two stages; i) for exact match and ii) for nearest match. This is done by two algorithms "Levenshtein Algorithm" and "Semantic Distance Algorithm".

N-ary Product Based Retrieval.
The translation for an input sentence is extracted in this stage. And there is possibility that input can have many translations. So the possibilities are collected and the idea of n-ary product is used to record all the feasible sentences.

Ordering of Translated Phrases.
If a single input sentence is divided into pieces and translated into output language phrase, then ordering of these translated phrases are done in this phase.

4 Comparison

4.1 Rule Based Machine Translation

The quality of translation in Rule Based Machine Translation (RBMT) depends upon large number of rules. Therefore its computational cost is very high. Rules are based on both source and target languages, their respective morphological, syntactical and

semantic structures. With a large set of large and fine grained linguistic rules, RBMT generates translation with acceptable quality, but developing system like this needs more time and man hours because this type of linguistic recourses should be hand crafted (Knowledge Acquisition Problem). As RBMT works with exact matches, it is unable to translate text when system does not have enough knowledge about the input. It is also difficult to add more rules for generating high quality output.

4.2 Statistical Machine Translation

The knowledge about translation is acquired automatically from the example data. This is the main reason why SMT is developed fast as compared to RBMT. In a situation where large corpus is available but linguistic knowledge is not readily available then SMT is a preferred method. When input and output languages are not complex morphologically then SMT techniques generate better results. SMT based approaches do not need Bilingual dictionaries. They depend upon the quality of bilingual corpus.

4.3 Example Based Machine Translation

It requires Bilingual dictionary. It translates text by adapting to examples. The computational cost is less than RBMT. By storing proper examples in the DB the system can be upgraded. It works on best matching reasoning, so therefore when the corresponding example is not available in corpus, the translation process becomes complicated. It translates in fail-safe way. Quality of translation depends upon the difference between input text and lookup results for similar examples. EBMT can also notify us that when its translation is improper.

Table 1. Comparison of RBMT, SMT and EBMT

	Advantages	Disadvantages
Rule Based Machine Translation	– Effective for core phenomena – Based on linguistic theories – Easy to build an initial system	– Rules are formulated by experts – Sometimes the experts do not agree hence the system remain unreliable. – Difficult to maintain and extend – Ineffective for marginal phenomena
Statistical Machine Translation	– Numerical knowledge – Extracts knowledge from corpus – Reduces the human cost – Model is mathematically grounded	– Less linguistic background – Overall lookup cost is high – Hard to capture long distance phenomena – Authenticity of results can be questionable. – Not suitable for free word order languages
Example Based Machine Translation	– Extracts knowledge from corpus – Based on translation patterns in corpus – Reduces the human cost	– Similarity measure is sensitive to system – Lookup cost can be high – Knowledge acquisition is problematic – Trade off is required between corpus size and performance.

5 Findings

The qualitative findings are tabulated in table 1, and the quantitative findings are mentioned in table 2. The BLEU metric is used for the evaluation of the machine translated text, five reference sentences were used for calculating the BLEU value. From the value of the BLEU it is clearly shown that EBMT performs better than the rest of the three systems. RBMT was found to be better than both the SMT systems. Out of the two SMT (Google and Bing), Bing translator gave better results than the Google translator.

Table 2. BLEU value of RBMT, EBMT and SMT

	RBMT	**EBMT**	**SMT**	
			Google	**Bing**
BLEU Value	0.8	**0.8421**	0.6268	0.709

6 Discussion

After detailed literature study and investigation of the above mentioned three MT systems, we can conclude that for languages with similar lexical and syntactic structure e.g. Urdu and Hindi, the Rule based MT technique gives better results. The SMT systems perform better if necessary resources such as annotated corpora etc. are available. At present, most of the systems translate text from source to target language on the basis of single sentence whereas in real life text for translation is much larger than one sentence. Nonetheless, the continuous process of repetitive translation and improvements by human annotators contribute significantly to any MT system.

7 Conclusion and Future Directions

In this paper we explained three main techniques of machine translation; Rule Based Machine Translation, Statistical Machine Translation and Example Based Machine Translation. We explained the methodology of each of these systems and found their comparison based on their respective outputs using carefully selected text. Our current work is preliminary in nature. However it reports significant results based on qualitative analysis.

In order to contribute a significant role to UMT research, at present we are in the process of building the required corpora. We intend to use our corpora to conduct larger scale automated experiments and report quantitative results that are comparable to human translators. Based on our qualitative and quantitative results, we aim at proposing a new model that minimizes flaws in the existing Urdu MT systems. Ideally, we would like to implement our proposed system with fewer requirements of computational and human resources.

8 REFERENCES

1. Abdullah, H., Homiedan.: Machine translation. J. King Saud Uni. Lang. & Trans. 10, 1-21 (1998)
2. Hutchins, J.: Latest Development in Machine Translation Technology: Beginning a New Era in MT RESEARCH. MT Summit IV, 11-34, Kobe, Japan (1993)
3. Lewis, Paul, M., Simons, G.F., Fennig, C.D.: Ethnologue: Language of the World. Seventeenth edition. Dallas, Texas: SIL International (2013)
4. Schmidt, R.L.: Urdu An Essential Grammar. Rutledge Taylor & Francis Group. London and New York (2004)
5. Ata, N., Jawaid, B., Kamran, A.: Rule Based English to Urdu Machine Translation. Proceedings of Conference on Language and Technology (CLT'07). (2007)
6. Jawaid, B., Zeman, D., Bojar, O.: Statistical Machine Translation between Languages with Significant Word Order Difference. Prague (2010)
7. Zafar, M., Masood, A.: Interactive English to Urdu Machine Translation using Example-Based Approach. IJCSE 1 (3), 276-283 (2009)
8. Ali, A., Siddiq, S., Malik, M.K.: Development of Parallel Corpus and English to Urdu Statistical Machine Translation. IJET-IJENS 10(5), 31-33 (2010)
9. Jawaid, B., Zeman, D.: Word-Order Issues in English-to-Urdu Statistical Machine Translation. Prague Bull. Math. Linguistics. 87-106 (2011)
10. Survey of Machine Translation Evaluation. EuroMatrix. (2007)
11. Samantaray, S.D.: Example based machine translation approach for Indian languages. ICCS. 1-10 (2004)
12. Somers, H. :Machine translation and Welsh: The way forward. A Report for the Welsh Language Board, Centre for Computational Linguistics, UMIST, Manchester (2004)