Statistical Machine Translation between Languages with Significant Word Order Difference

by

Bushra Jawaid

M.Sc. Thesis

European Masters Program in Language and Communication Technology (LCT)

University of Malta
Dept Intelligent Computer Systems
&
Charles University in Prague
Faculty of Mathematics and Physics

Supervisor: RNDr. Daniel Zeman, Ph.D.
Mr. Mike Rosner

Prague, 2010

[This page is intentionally left blank]

By taking this opportunity I would like to acknowledge my supervisor Daniel Zeman for his endless support, encouragement, helpful comments and corrections on my grammatical mistakes.

I am also thankful to all my coordinators and teachers who paved the way for this Masters Program and special thanks to Ondrej Bojar for his comments on this study work and help in running experiments. A special acknowledgement to Mr. Tafseer Ahmed who is the inspiration and motivation behind all my work in NLP. And finally I would like to thank my family and friends for their support throughout my studies.

I certify that this master thesis is all my own work, and that I used only cited literature. I agree with making this thesis publicly available.

Prague, August 05, 2010 Bushra Jawaid

[This page is intentionally left blank]

Table of Contents

LIST OF FIGURES

Title: Statistical Machine Translation between Languages with Significant Word Order Differences
Author: Bushra Jawaid
Department: Institute of Formal and Applied Linguistics
Supervisor: RNDr. Daniel Zeman, PhD.
Supervisor's email address: zeman@ufal.mff.cuni.cz

Abstract:
One of the difficulties statistical machine translation (SMT) systems face are differences in word order. When translating from a language with rather fixed SVO word order, such as English, to a language where the preferred word order is dramatically different (such as the SOV order of Urdu, Hindi, Korean, ...), the system has to learn long-distance reordering of the words. Higher degree of freedom of the word order of the target language is usually accompanied by higher morphological diversity, i.e. word affixes have to be generated based on the fixed word order in the source sentence.

The goal of the thesis is to explore the two mentioned (and possibly other related) classes of problems in practice, and to implement and evaluate techniques expected to help the SMT system to solve them. This includes:

1. Selecting a language pair with word order differences and collecting parallel data for the pair.

2. Training an existing SMT system on the data.

3. Evaluating the performance of the system and analyzing the errors it does. Estimating how much the accuracy of translation is affected by the problems mentioned above, and possibly what are the other types of error causes that dominate the output.

4. Implementing preprocessing and/or other techniques aimed at minimizing the found classes of errors. Evaluating their impact.

Keywords: Statistical Machine Translation, syntactic word order differences, rich morphological languages, parallel corpus

[This page is intentionally left blank]

Chapter

1

Introduction

Natural Language Processing (NLP) is a branch of Artificial Intelligence devoted to the study of computerized approach to analyze, generate and represent the human language[1]. The representation of human language is defined on certain levels of linguistic analysis for achieving the human-like processing. From linguistic point of view, these levels of dependencies are: morphology, syntax, semantic and pragmatic (Jurafsky, et al., 2000). Each level in NLP is highly ambiguous[2] when it comes to computationally model the language. Thus, the goal of NLP is *to accomplish unambiguous human-like language processing*. To achieve this goal we need to build computer systems that can translate the text from one language into another, answers the queries about the content of the text and is able to draw inferences from the text.

For several decades dating back to the late 1940s, NLP has been one of the most active areas of research. Machine Translation (MT) was the first computer-based application developed under the field of NLP. The task of an MT system is to translate the text or speech from one language into text or speech in another language. There are many approaches to MT that are roughly classified in two paradigms: Rule-based and Data-driven.

In a classical rule-base system deep analysis of linguistic phenomenon of the given language pair is performed. Rule-base systems usually consist of a set of transformation rules written by human expert and an MT engine, where linguistic knowledge is represented through that set of rules. Rule-based system involves three phases: analysis, transfer, and generation. Source sentence is analyzed using parsers and/or morphological tools, gets transformed into intermediate representation using the transfer rules, and then target language sentence is generated from the intermediate representation.

In the Data-driven approach large text corpora are used to develop the approximated generalized models of linguistic phenomena based on the actual examples of these phenomena that are provided by the text corpora without adding any significant linguistic or world knowledge. The data driven approach has the advantage over the possibility of using the

[1] *http://www.pcai.com/web/ai_info/natural_lang_proc.html*

[2] *http://www.site.uottawa.ca/tanka/files/complexities.html*

same system for translating any pair of languages for which enough training data is available. The further classification of the data driven approach is made between the example-based approach, where the basic idea is to do translation by analogy and the statistical approach. In statistical approach, Bayes decision rule and statistical decision theory are used to minimize the number of errors to get the best translation from source language to target language.

The Statistical approach has several advantages over the other translation schemes. Often the relationships between linguistic objects such as words, phrases or grammatical structures are difficult to model but, in statistical translation systems these dependencies can be automatically learnt from the training data. As model parameters are learnt from training data, adding more and more data into the system makes it better.

Among the different approaches to Machine Translation described above, our main focus in this study is *Statistical Machine Translation*[3] (SMT). This thesis primarily focuses on ***English to Urdu Statistical Machine Translation System***. The selection of this language pair is due to the linguistic characteristics each language hold related to our task. The goal of this study is to achieve the improvement in translation quality for the given language pair by using the linguistic knowledge of either source or target or both languages.

The rest of the chapter continues with the English and Urdu languages specification together with the morphological and syntactic differences in both languages. Then we give the brief overview of statistical machine translation systems and the recent work in the field of English to Urdu SMT. After introducing the issues in modeling the selected language pair and the architecture of SMT systems, we define our goals for this study.

1.1 Source and Target Language Features

As we already mentioned above, for this study we have selected English as the source and Urdu as the target language for the translation purpose. English is read and written from left-to-right whereas Urdu is read and written from right-to-left. Both languages differ in morphological and syntactic features; English has a relatively simple inflectional system, only nouns, verbs and sometimes adjectives can be inflected, and the number of possible inflectional affixes is quite small (Jurafsky, et al., 2000). Urdu on the other hand is highly inflectional and rich in morphology. In Urdu verbs and adjectives are inflected according to gender, number and person of the head noun and noun phrases inflect according to their gender, number and case.

English is a fixed word-order language and follows the S-V-O (Subject-Verb-Object) structure; whereas Urdu is a free word-order language and

[3] *http://www.statmt.org/*

allows many possible word orderings but, the most common sentence structure used by the native speakers is S-O-V. The other major difference is the existence of prepositional part-of-speech in English whereas Urdu noun and verbs are followed by postpositions. Both languages are linguistically different from each other and thus translation between both languages is not very straight forward.

For the readers who are not familiar with the Urdu language we provide the basic Urdu alphabetical set in Table 1.1 with the Unicode values and IPAs (International Phonetic Alphabet). Figure 1.1 shows the representation of each cell in Table 1.1. Alphabets are positioned vertically from top left corner.

ا (ɑ) 0627	ث (s) 062B	د (d) 062F	ز (z) 0632	ض (z) 0636	ف (f) 0641	م (m) 0645	ء (I) 0621
ب (b) 0628	ج (ʤ) 062C	ڈ (ɖ) 0688	ژ (ʒ) 0698	ط (t̪) 0637	ق (q) 0642	ن (n) 0646	ی (j) 06CC
پ (p) 067E	چ (ʧ) 0686	ذ (z) 0630	س (s) 0633	ظ (z) 0638	ک (k) 06A9	و (v) 0648	ے (e) 06D2
ت (t̪) 062A	ح (h) 062D	ر (r) 0631	ش (ʃ) 0634	ع (ʔ) 0639	گ (g) 06AF	ہ (h) 06C1	
ٹ (ʈ) 0679	خ (x) 062E	ڑ (ɽ) 0691	ص (s) 0635	غ (ɣ) 063A	ل (l) 0644	ھ (h) 06BE	

Table 1.1: Urdu Alphabet Chart with IPA and Unicode

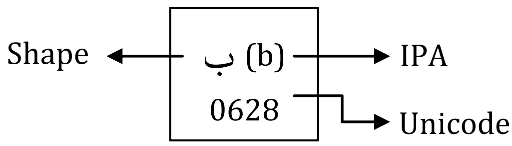


Figure 1.1: Cell representation of Table 1.1

We also provide the small example of English and Urdu parallel sentence pair with the Word-to-Word gloss in Example 1.1.

Example 1.1:

English Sentence:	Do you understand English and Urdu?						
Urdu Translation:	؟	سمجھتے ہیں	اُردُو	اور	انگریزی	آپ	کیا
Transliteration:	?	samjhte heñ	urdū	aor	angrezī	āp	kyā
Gloss:	?	understand	Urdu	and	English	you	do

1.2 Overview of Statistical Machine Translation System

Statistical machine translation system is one of the applications of Noisy Channel Model introduced by (Shannon, 1948) using the information theory. The goal of the probabilistic noisy channel model can be summarized as:

What is the most likely sentence out of all sentences in the language E given some input in foreign Language F?

As illustrated in Figure 1.2, the setup the noisy channel model of a statistical machine translation system for translating from Language F to Language E works like this: The channel receives the input sentence e of Language E from source, transforms it into a sentence f of language F and sends the sentence f to a decoder. The decoder then determines the sentence $\hat{e}$ of language E that f is most likely to have arisen from and which is not necessarily identical to e.

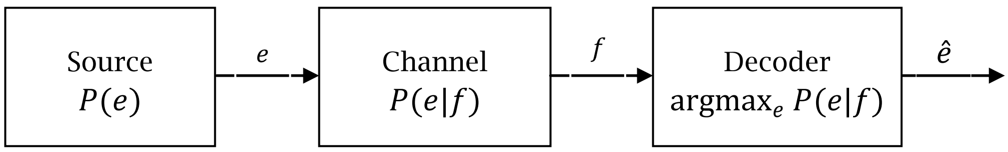


Figure 1.2: The noisy channel model of statistical machine translation system.

Thus, for translation from language F to language E, statistical machine translation system requires three major components. A component for computing probabilities to generate sentence e, another component for computing translation probabilities of sentence f given e, and finally, a component for searching among possible foreign sentences f for the one that gives the maximum value for $P(f|e)P(e)$.

Note: In this introductory chapter notion $P(.)$ is used to show general probability distribution with almost no specific assumption, while $p(.)$ is used for model-based probability distribution.

Let's treat each sentence as composition of string of words. Assume that a sentence f of language F, represented as $f_1^J = f_1, \dots, f_j, \dots, f_J$, is translated into a sentence e of language E, and represented as $e_1^I = e_1, \dots, e_i, \dots, e_I$. Then, the probability, $P(e_1^I|f_1^J)$ assigned to a pair of sentences $\left(f_1^J, e_1^I\right)$, is interpreted as the probability that a decoder will produce the output sentence e_1^I given the source sentence f_1^J.

$$\hat{e}_1^I = \underset{e_1^I}{\operatorname{argmax}} \{ P(e_1^I|f_1^J) \} \tag{1.1}$$

Equation 1.1 is also known as Bayes Decision Rule. For translating sentence f_1^J into sentence e_1^I, we need to compute $P(e_1^I|f_1^J)$. For any given probability $P(y|x)$, it can be further broken down using Bayes' theorem.

$$P(e_1^I|f_1^J) = \frac{P\left(f_1^J \middle| e_1^I\right) . P(e_1^I)}{P(f_1^J)} \qquad 1.2$$

Since we are maximizing over all possible sentences for the given sentence f_1^J, Equation 1.2 will be calculated for each sentence in Language E. But $P(f_1^J)$ doesn't change for each sentence. So we can omit the denominator $P(f_1^J)$ from the Equation 1.2.

$$\hat{e}_1^I = \underset{e_1^I}{\operatorname{argmax}} \left\{ P\left(f_1^J \middle| e_1^I\right) . P(e_1^I) \right\} \qquad 1.3$$

Now consider the first term in Equation 1.3, $P\left(f_1^J \middle| e_1^I\right)$ likelihood of translation (f, e) is called **Translation Model** and the second term $P(e_1^I)$ the prior probability is called **Language Model**.

1.2.1 Language Model

A Language Model (LM) is a probability distribution over the possible strings (which we can represent either as $w_1 \dots w_n$ or w_1^n) of a language that attempts to reflect how frequently a string of words w_1^n, occurs as a sentence. Depending on the language, a Language Model can be defined over sequences (word or Part-of-Speech sequences) or over structures (utterance-tree pairs). In this section we describe the n-gram language model over sequence of words. Where, in n-gram model the task of predicting the next word can be stated as attempting to estimate the probability function P (Manning, et al., 1999).

$$P(w_n|w_1 , \dots , w_{n-1}) \qquad 1.4$$

If we consider each word occurring at a specific position in a sequence of string is an independent event then the probability over sequence of words is $P\ (w_1,\ w_2, \dots, w_{n-1}, w_n)$ or $P(w_1^n)$ (Jurafsky, et al., 2000). Using the chain rule of probability we can decompose this probability:

$$\begin{aligned} P(w_1^n) &= P(w_1) .\ P(w_2|w_1) . P(w_3|w_1^2) \dots P(w_n|w_1^{n-1}) \\ &= \prod_{i=1}^{n} P\left(w_i \middle| w_1^{i-1}\right) \end{aligned} \qquad 1.5$$

Hence, the probability of word sequence is calculated by conditioning the next word on the history seen so far. But for instance, to compute the probability of a word w_n given a long sequence of preceding words is not a trivial task. To solve this problem model is usually approximated by applying Markov assumption. According to Markov assumption only the prior local context, consisting of last few words, affects the next word. Thus, in Markov models the probability of the next word depends only on the previous k words in the word sequence. In general, an N-gram is an $(N-1)th$ order Markov Model. For instance, Markov model with k = 1 is called bigram model because it depends on one previous word only:

$$P(w_1^n) \approx \prod_{i=1}^{n} P(w_i|w_{i-1}) \qquad 1.6$$

We need a large monolingual training corpus of flat sentences to train language model. In order to build the bigram language model the probability $P(w_i|w_{i-1})$ can simply be estimated by counting the frequencies of the event $\langle w_{i-1}\, ,\ w_i \rangle$. This technique of probability estimation is called the Maximum Likelihood Estimate (MLE), shown in Equation 1.7.

$$P(w_i|w_{i-1}) = \frac{count\ (w_{i-1}\, ,\ w_i)}{\sum_w count\ (w_{i-1}\, ,\ w)} = \frac{count\ (w_{i-1}\, ,\ w_i)}{count\ (\ w_{i-1})} \qquad 1.7$$

But the MLE is in general unsuitable for statistical inference in NLP. The problem is the sparseness of our data. The MLE assigns a zero probability to unseen events, and since the probability of a long string is generally computed by multiplying the probabilities of subparts, these zeros will propagate and give us bad (zero probability) estimates for the probability of sentences when we just happened not to see certain n-grams in the training text (Manning, et al., 1999). To overcome this problem a technique called smoothing is used. Smoothing works by decreasing the probability of previously seen events, and assigns the leftover probability mass to previously unseen events. There are number of smoothing methods available, like adding 1 to the counts, Good Turning estimates, smoothing using general linear interpolation etc. (Chen, et al., 1998)presents detailed discussion on different smoothing algorithms.

Although training the lower order language model causes loss of information because of limited history, even then usually uni-, bi-, or trigram language models are used. Actually, training the high order language model again reveals the data sparseness problem. Still, existence of a language model is very crucial in SMT, it helps in selecting the fluent translation for the given source sentence.

1.2.2 Translation Model

For translating the string translation probability $P(f_1^J|e_1^I)$ (in Equation 1.2), different translation models schemes have developed in the field of SMT till date, based on encountered language dependent issues. The most well known translation schemes are: word-base translation, phrase-based translation, and tree-based translation.

Single-Word-based Translation Models

The basic idea of single-word based approach is to segment the given source sentence into words, then translate each word and finally compose the target sentence from word translations. The key issue in modeling the string translation probability is to identify the correspondence between the words of the source sentence and the words of the target sentence. Let's assume all word pairs (f_j, e_i) of the given sentence $(f_1^J;\ e_1^I)$ that have sort of pairwise dependence, the models describing these type of dependencies are known as Alignment Models.

Word Alignments

There are two general approaches to word alignments: statistical models and heuristics models. In this section we briefly discuss the basic statistical alignment model.

To model the translation probability $P(f_1^J|e_1^I)$, word alignment $a_1^J := a_1 \dots a_j \dots a_J$ is introduced in the translation model as the hidden variable, which describes the mapping from source position j to a target position a_j. The relationship between alignment model and translation models is given by:

$$P(f_1^J|e_1^I) = \sum_{a_1^J} P(f_1^J, a_1^J|e_1^I) \qquad 1.8$$

There are different decompositions of the probability distribution $P(f_1^J, a_1^J|e_1^I)$ based on the statistical models. Here we are discussing the basic alignment model (Zens, et al., 2002) decomposition approach. By applying the chain rule, model is further factorized as:

$$P(f_1^J|e_1^I) = \sum_{a_1^J} P(a_1^J|e_1^I) \quad P(f_1^J|a_1^J, e_1^I) \qquad 1.9$$

$$= P(J|e_1^I) \sum_{a_1^J} \prod_{j=1}^{J} [p(a_j|a_{j-1}, I, J).p(f_j|e_{aj})] \qquad 1.10$$

Here, we have the following probability distributions:

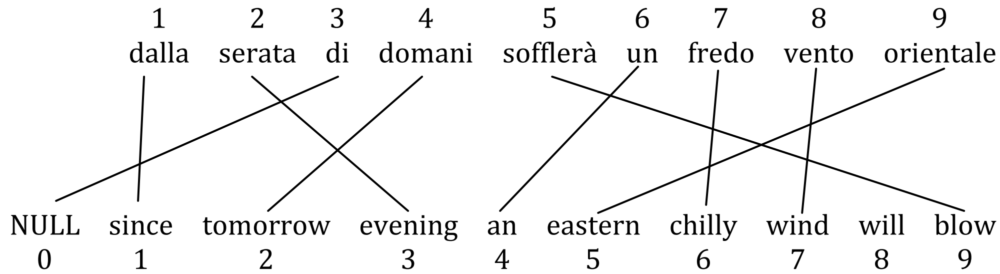
$P(J|e_1^I)$ = the sentence length distribution, which is included in the formula for completeness but can be omitted without any loss of performance.

$p(f|e)$ = the lexicon probability.

$p\left(a_j|a_{j-1}, I, J\right)$ = the alignment probability.

In Equation 1.10, a_j is the position in e_1^I that f_j is aligned with; e_{aj} is the word in e_1^I with that f_j is aligned. The basic idea of the formula showed in Equation 1.10 is J times summing over all possible alignments of source sentence to target sentence. The meaning of $a_j = 0$ for position a_j is null alignment of word in source sentence at position j with any word in target sentence that means it has no obvious translation. According to the formula explained in Equation 1.10, each target word can be mapped on more than one word in source sentence but many-to-one alignment from source to target is not allowed. During word alignment word reordering can also be performed.

Example 1.2:



In Example 1.2 (Federico, 2009) we can see the word alignment in an Italian-to-English sentence pair. In this example we see the possible alignments explained earlier: null alignment of word “di” from source to target, many-to-one alignment and also word reordering induced by alignments.

To compute the probability of the alignment (dalla serata di domani sofflerà un fredo vento orientale | NULL(3) since(1) tomorrow(4) evening(2) an(6) eastern(9) chilly(7) wind(8) will blow(5)), we multiply the 9 (length of source sentence) translation probabilities. The probability of this alignment is calculated as:

P(dalla|since) * P(serata|evening) * P(di|NULL) * P(domani|tomorrow) * P(sofflerà|blow) * P(un|an) * P(fredo|chilly) * P(vento|wind) * P(orientale|eastern)

There are various ways to model the translation probability. The most popular statistical translation models are IBM-1 to IBM-5 (Brown, et al., 1993) and Hidden-Markov alignment model (HMM). These models are discussed in detail in (Och, et al., 2000). These models differ in translation models but the lexicon probability $p(f|e)$ is based on single words in both source and target languages. Brief introduction of all these models is as follows:

- In IBM-1 uniform distribution, $p(i \mid j, I, J) = 1/(I+1)$, is used i.e. all alignments have same probability.

- IBM-2 adds the absolute reordering model. It is based on zero-order alignment model $p(a_j | j, I, J)$ where different alignment positions are independent from each other.

- The HMM models use the first-order model where to reduce the number of parameters, the dependence on J is ignored and distribution $p(a_j | a_{j-1}, I)$ is used instead of $p(a_j | a_{j-1}, I, J)$. In this distribution a_j depends on the previous alignment a_{j-1}.

- In IBM-3, we have an (inverted) zero-order alignment model $p(j | a_j, I, J)$ with an additional fertility model $p(\Phi|e)$which describes the number of words Φ aligned to an English (target) word e (Zens, 2008).

- IBM-4 adds the relative reordering model. It is based on (inverted) first-order alignment $p(j|j', I, J)$ and fertility model $p(\Phi|e)$ (Zens, 2008).

- IBM models have some serious draw-backs. These models don't allow many-to-one alignment mapping from source to target, i.e. target word can be aligned with at most one foreign word. To resolve this issue some transformations can be applied; Parallel corpus aligns in both directions and word alignment from source to target and target to source are generated. The union of both directional alignment points provides high-recall alignment with additional alignment points whereas taking the intersection of both alignments gives the high-precision alignment with high-confidence alignment points.

Phrase-Based Translation Models

The main disadvantage of the word-based translation systems over phrase-based translation (PBT) models is that in single-word based (SWB) approach contextual information is not taken into account. In languages, many linguistic phenomena have more than single-word dependencies. "For many words, the translation depends heavily on the surrounding words. In the SWB translation, this disambiguation is done completely by the language model. Often the language model is not capable of doing this. An example is shown in Example 1.3" (Zens, et al., 2002).

Example 1.3:

Source:	Was halten Sie vom Hotel Gewandhaus?
Target:	What do you think about the hotel Gewandhaus?
SWB:	What do you from the hotel Gewandhaus?
PBT:	What do you think of hotel Gewandhaus?

The translation from German to English in Example 1.3 shows the influence of neighboring words on the translation. In languages, translation of compound words, literal translations and many other phenomena are problematic for single-word alignment. In PBT many-to-many translations can be learned and also huge training data helps in learning longer phrases and results in better translation. PBT also supports translation of non-compositional phrases i.e. phrases whose meaning is determined by taking the collective meaning of all components of phrases instead of their individual meanings (like real estate, face value) .

In PBT models, a phrase is merely considered as sequence of words. The context is included in the phrase translation models by considering the chunk of words (phrases) instead of single words. In phrase-based approach as opposed to single-word approach, the source is segmented into number of phrases, each phrase is translated independently and finally the target sentence is formed by combining all those phrase translations.

Approaches for learning Phrase-Based Translation

Different approaches have been introduced to learn phrase based translations. Most of these approaches are based on word alignments whereas (Marcu, et al., 2002) propose to establish lexical correspondence at the phrase level instead of word level. To learn such correspondences, they introduced a phrase-based joint probability model that simultaneously generates both the source and target sentences in the parallel corpus.

(Koehn, et al., 2003) presented the phrase model based on the word alignments. They collect all word pairs that are consistent with the word alignment and the phrase alignment of those word pairs contains all the alignment points for all the words it covers. Then, for all the collected phrase pairs, phrase translation probability is estimated using the relative frequency. Reordering of the target output phrases is modeled through relative distortion probability distribution $d(start_i, end_{i-1})$, where $start_i$ refers to the starting position of foreign phrase that is translated into ith target phrase, and end_{i-1} refer to the end position of the foreign phrase that is translated into $(i-1)th$ target phrase. The simple distortion model with suitable α value is used:

$$d(start_i, end_{i-1}) = \alpha^{|start_i - end_{i-1} - 1|} \quad 1.11$$

The translation probability is calculated as:

$$P(f_1^J | e_1^I) = \Phi_{i=1}^{I} \, \phi(f_i, e_i) \; d(start_i, end_{i-1}) \quad 1.12$$

Where,

Each f_i and e_i represents the foreign phrase and target phrase respectively.

$\phi(f_i, e_i)$ = probability distribution.

(Och, et al., 1999) presented the alignment template approach due to deficiency in baseline alignment models. Baseline models can only create the correspondence between single words. In this approach word classes are used instead of words and, alignment templates are used to generalize the phrases. The alignment template is defined as the triple $z = (\tilde{F}, \tilde{E}, \tilde{A})$ where $\tilde{A}$ refers to the alignment between source class sequence $\tilde{F}$ and a target class sequence $\tilde{E}$.

If we have to calculate the translation probability of (bruja verde| green witch), then the Figure 1.3 (Knight, et al., 2004) shows the alignment template that covers the source sentence and the produced translations.

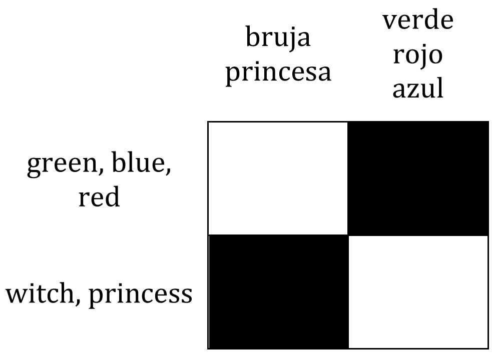


Figure 1.3: Alignment Template approach

Even after applying further improvements in the word alignments, phrase model suffers from the problem of modeling non-contiguous phrases i.e. phrases with the gap in the middle. Also, phrase-based translations cannot deal with Syntactic transformations during translation because PBT don't account for linguistic features.

Summary

In this section we briefly discussed the working pipeline and components of the translation system based on noisy channel model. In Figure 1.4 we illustrate the complete architecture of Statistical Machine Translation.

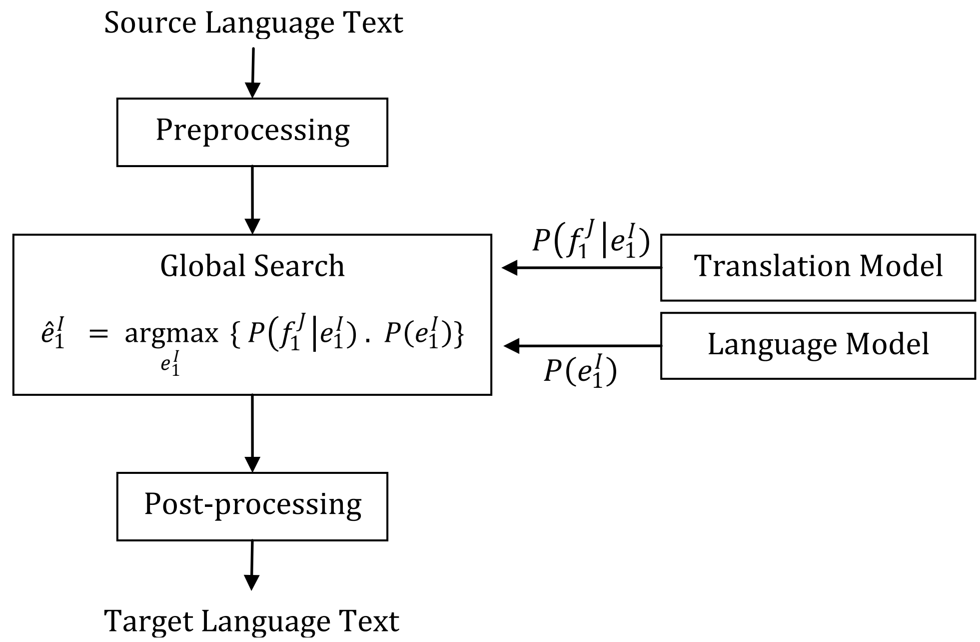


Figure 1.4: Architecture of translation approach based on Bayes Decision Rule

Tree-Based Translation Models

There are some major issues with the PBT models:

- PBT systems are (mostly) based on IBM word alignment models and IBM translation models don't model structural or syntactic aspect of language. These models are well suited for the structurally similar language pairs like English and French. The language pairs that differ in word order cannot be well modeled using PBT models.
- Another issue which PBT systems face is data sparseness. This problem becomes more complicated for highly inflectional languages.
- PBT systems have also introduced reordering options but still they are unable to deal with global reordering because the distortion model is based on movement distance (distance-

based model gives linear cost to the reordering distance) that may face computational resource limitations (Och, et al., 2004).

Hence, all the above mentioned problems give rise to the introduction of tree-based models in the field of SMT. For tree-based models decoding is not linear with respect to sentence length, unless reordering limits are used. Tree-based models use both linguistically sound syntax-based models i.e. models that have non-terminals based on syntactic categories (noun phrase (NP), verb phrase (VP) and so on), and formally syntax-based models i.e. those based on single non-terminal (X).

There are a few terms used in this field that should be clear before going through the rest of the section. The common terms used in the literature to represent tree-based models that are introducing the syntax are: hierarchical phrase-based, tree-to-string, string-to-tree, syntax-augmented, syntax-directed, syntax-based and others.

Hierarchical phrase-based systems don't use real linguistic syntax while syntax-directed and syntax-augmented use linguistic syntax only in the source language and target language respectively. The other models, string-to-tree and tree-to-string use the linguistic syntax only in the target language and source language respectively. Syntax-based models can either be build using syntax trees generated by parsers or using tree transfer methods motivated by syntactic reordering patterns.

Formalism

Formalism for hierarchical phrase-based and syntax-augmented is Probabilistic Synchronous Context-Free Grammar (PSCFG), the PSCFG translation models define weighted transduction rules that are defined as source and target terminal sets and a non-terminal set:

$$X \rightarrow \langle \alpha, \beta, \sim, \omega \rangle \qquad 1.13$$

Where X is non-terminal, α is a set of source language terminals and non-terminal, β is a set of target language terminals and non-terminals, $\sim$ is a one to one mapping from set of non-terminals in α to set of non-terminals in β and ω represents the non-negative weight assigned to each rule.

Translation with a PSCFG is thus a process of composing such rules to parse the source language while synchronously generating target language output (Zollmann, et al., 2008). The PSCFG rules are automatically learned from parallel training data. These rules capture the syntactic ordering of the words in the language and by using non-terminal symbols/categories generalize beyond the lexical level.

The hierarchical phrase-based models combine the insight of the phrase-based models with syntactic structures. The use of a hierarchical model was first presented by (Chiang, 2005). In his model, hierarchical phrases are used instead of simple phrases, where hierarchical phrases are composed of words and sub phrases. In the proposed translation model he used synchronized CFG together with the “glue” markers. The PSCFG rules are learned using the bilingual phrase pairs of phrase-based MT. The gluing rules are used to combine the sequence of X to form sentence S. Example 1.4 (Chiang, 2005) shows how the hierarchical phrase pairs from Chinese to English are formalized in a synchronous CFG:

Example 1.4:

PSCFG Rules:

$$X \rightarrow \langle \text{ yu } X_1 \text{ you } X_2 \text{ , have } X_2 \text{ with } X_1 \rangle$$

$$X \rightarrow \langle X_1 \text{ de } X_2 \text{ , the } X_2 \text{ that } X_1 \rangle$$

Glue Rules:

$$S \rightarrow \langle S_1 X_2 , S_1 X_2 \rangle$$

$$S \rightarrow \langle X_1 , X_1 \rangle$$

Where, subscripts are used to indicate the reordering of the phrases defined as mapping set ~ in Equation 1.13.

In rest of the tree-based models other than hierarchical models syntactic parsers are used to get parse tree of source language, or target language, or both. (Yamada, et al., 2001) used approach of tree-to-string based translation models and (Eisner, 2003) presented translation model based on non-isomorphic tree- to-tree mapping. Yamada used (Collins, 1999) parser to parse source side (English) of the corpus. After getting the parse tree they perform operation on each node of the tree. The operations are: reordering child nodes, inserting extra words at each node, and translating leaf nodes. The example of the operations they performed to get transformed tree are shown in Figure 1.5.

Another interesting research filed towards syntax-based machine translation is dependency-based translation. In this approach translation is performed using dependency structures instead of using Context free grammars. Work based on similar approach for Czech-English is presented by (Čmejrek, et al., 2003). (Zollmann, et al., 2008) and (Khalilov, et al., 2009) have further provided a brief introduction and comparisons among phrase-based, hierarchical and syntax-augmented models.

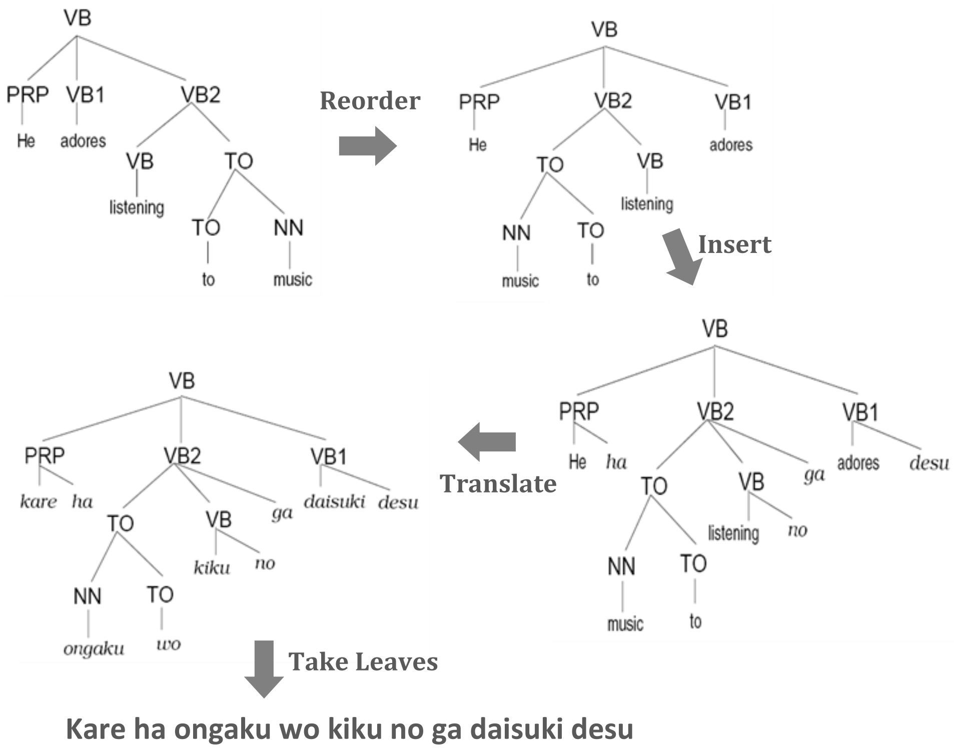


Figure 1.5: Yamada's translation operations: Reorder, Insert, Translate

1.2.3 Decoder

The goal of the decoder is to take the model, estimate the parameters of the model and to perform the actual translation. The translation tables are the main knowledge source for the machine translation decoder. The decoder consults these tables to figure out how to translate input in one language into output in another language. The process of decoding corresponds to maximizing the Bayes decision rule defined in Equation 1.1. Optimizing the maximization function in decoding process is quite difficult task because a huge search space of possible candidate target language sentences is to be considered for a given input sentence. Therefore, a primary function of decoder is to search this space as efficiently as possible. (Lopez, 2008) has categorized decoders into two main categories: FST Decoders and SCFG Decoders.

Decoders under these categories also provide search techniques that sacrifice optimality over efficiency. A brief introduction of A*-based stack search techniques is presented in (Brown, et al., 1990). Read (Lopez, 2008) for further detail on types of decoder, working knowledge of decoder and further references.

1.3 Our goals

- **Collection of Parallel and Monolingual Corpora:** The bilingual and monolingual corpuses are the starting point for statistical machine translation. For this study we collect the English to Urdu parallel corpora and also large monolingual Urdu corpus for the Language Model.

- **Data Reordering:** Both English and Urdu languages differ in word order, and translation between languages with different word order is not trivial task. In this study we try to improve the translation quality of the system by pre-processing the source side of the parallel corpora. We use the transformation scheme to change the word order of the English sentence according to the default sentence structure of Urdu.

- **Factorization:** To overcome the data sparseness issue we use the factorized model of the phrase-based MT

1.4 Related work

Recently Google added the English to Urdu (Alpha) Statistical Machine translation system[4] in its 19th stage of research work. Google is based on Statistical machine translation approach and their research is inspired by the research work of Franz-Josef Och. Google has its own translation system for translating language pairs. The system is Alpha released so not all the technical details are publicly available yet. Also, we couldn't gather the information about the parallel corpus used for the translation. As Google mostly uses the million words corpus for the translation between languages, so they might have also collected the huge bilingual corpus for English and Urdu. The translation output for English to Urdu shows that they have relatively high amount of news data in their bilingual corpus. We have compared some of the translation output results from Google with the results produced by our system in this study work. The one important observation in Google's translation output is, currently they are not using their English to Urdu transliteration system for untranslated words. With the use of their transliteration system for untranslated words they might improve the translation quality of the Alpha system.

1.5 Outline of the thesis

Chapter 2 starts with the collection of parallel and monolingual corpuses for this study and also the detail of all the methods and techniques used to collect corpora. This chapter also presents the statistics over the collected

[4] *http://translate.google.com/#en|ur|*

corpora and the normalization techniques performed on the corpora for the improvement in translation quality.

Chapter 3 introduces the translation system used for this study and the issues associated with the selected MT system. To overcome those issues we explore the language dependent methodologies for the improvement in translation quality.

Chapter 4 comprises the experimental setup and the different range of experiments performed during the study. Error analysis is being done on the output of all experiments and the results are compared in terms of translation quality and the evaluation measure used for this study.

Chapter 5 concludes the overall study by summarizing the results and drawing conclusions on the basis of the results achieved after applying the techniques to improve the machine translation output. Further this chapter concludes by giving the suggestions for the improvement of the results attained in this study.

Chapter 2

Corpus Collection

Statistical machine translation systems always need good quality sentence by sentence aligned parallel data for the system training. The good quality parallel data helps in producing better quality translation results. But the most important part is the amount of the parallel data in hand, more parallel data ensures that the output translation will be more human understandable. Besides the parallel bilingual corpus we also need a large monolingual corpus in target language. The monolingual corpus is used to build a language model that helps to make the translation more fluent. The main concern for this study was the unavailability of English-to-Urdu ready-to-use parallel corpus. To begin with this study work we were provided with two parallel corpora from diverse domains. We collected rest of the bilingual corpora and entire monolingual corpus by web crawling.

Below is the statistics of the data collected for this research study and also the discussion on the problems faced during the searching for resources of parallel data. We also applied normalization on the data after finishing the collection phase to make it usable for the training of translation system.

2.1 Collection of Bilingual Data

For this study four different parallel corpora of at least three different domains were collected from various sources. The description of data collection, data resources and data processing is discussed in detail in this section.

2.1.1 Emille Corpus

For the bilingual corpus collection our first motive was to collect data from as different domains as possible to get better translation quality and a wide range vocabulary. For this purpose the first corpus we selected to use in our study is Emille (Enabling Minority Language Engineering). EMILLE is a 63 million word corpus of Indic languages (Baker, et al., LREC' 2002) which is distributed by the European Language Resources Association (ELRA).

Emille contains data from six different categories: consumer, education, housing, health, legal and social. This data is based on the information leaflets provided by the UK government and the various local authorities. We were provided in total 72 parallel files with each filename consisting of language code, text type (written or spoken), genre and subcategory, connected with hyphen character. The data is encoded in full 2-byte Unicode format and marked up in SGML format. The further detail about Emille corpus is available from their online manual[1]. The approach of the data extraction and processing on data is illustrated in Figure 2.1, and described below.

i. SGML to text: the sentences are extracted from the SGML tagged data using the program written in .net. The structure of each Emille document is as follows: it consists of header information and main text. Inside main text we have paragraphs and each paragraph consists of multiple sentences. We extract all the sentences from each paragraph and store them on the disk. The result of this phase is unaligned parallel sentences.

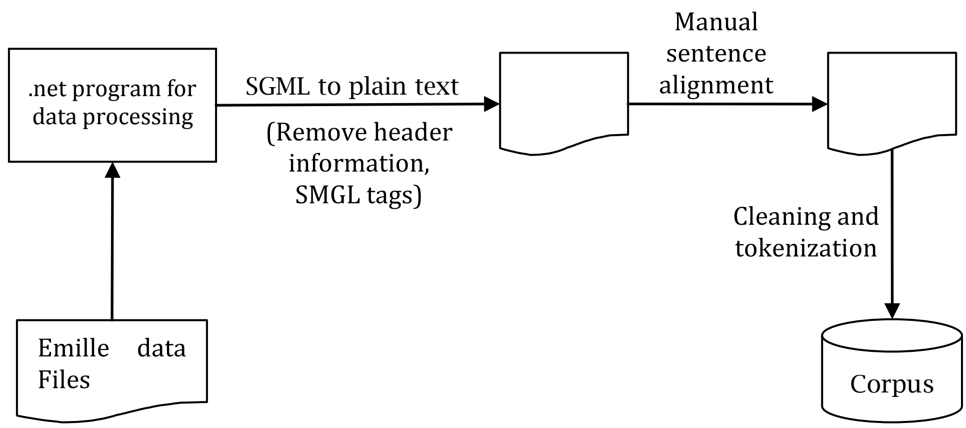


Figure 2.1: Overview of Corpus Creation from Emille Corpus

ii. Sentence alignment: we have manually aligned the sentence pairs that are extracted from the marked up text. The details of issues in corpus and manual sentence alignment are discussed in Section 2.3.

iii. Cleaning and tokenization: firstly, we clean the corpus before performing tokenization. Cleaning of corpus includes: removing blank lines from the data and removing bad characters from the data. As our tokenization script doesn't

[1] *http://www.lancs.ac.uk/fass/projects/corpus/emille/MANUAL.htm*

delete the blank lines so we do it in cleaning step. For removing bad characters, the analysis was performed once and the Unicode of all those characters are listed down that are not part of the text. Text is then cleaned from all those unwanted characters that are listed during the analysis phase. After cleaning the corpus we tokenize the cleaned data; data tokenization is discussed in the Section 4.1.1 in experimental setup.

2.1.2 Penn Treebank Corpus

Penn Treebank corpus (Marcus, et al., 1999) is the 2nd next wide domain corpus we have picked for this study. All the Penn Treebank data is released through the Linguistic Data Consortium (LDC). The parallel Penn-Urdu[2] Treebank data is released by the Centre for Research in Urdu Language Processing (CRULP) under the Creative Common License[3]. The corpus is freely available online[4] for the research purpose. The translation of Penn-Urdu Treebank is just a plain text and it is not available in Treebank format anymore. Also the whole Treebank-3's translation in Urdu is not yet available, only subpart of the Penn Treebank is used in this work.

Penn Treebank-3 is a bank of linguistic trees where each parse tree contains the syntactic and semantic information. Trees are annotated with part-of-speech-tag and special bracketing style is used for the extraction of predicate-argument structure. Penn Treebank is the collection of Wall Street Journal (WSJ), Brown corpus, Switchboard and ATIS. In this work we have only used the collection of WSJ stories that are distributed in both Penn Treebank-2 and Treebank-3. The Penn Treebank contains 2,499 stories from WSJ and they are distributed in 25 folders with 100 stories in each folder. For this study we have used only 317 stories whose Urdu translation is also available. The detail of used WSJ sections[5] is provided by the CRULP. For the collection of corpora from Penn-English Treebank the same procedure is used as described above but with few differences that is illustrated in Figure 2.2 and described below.

---

[2] *The work has been supported by the Language Resource Association (GSK) of Japan and International Development Research Center (IDRC) of Canada, through PAN Localization project (www.PANL10n.net).*

[3] *http://www.crulp.org/software/license/CreativeCommons.html*

[4] *http://crulp.org/software/ling_resources/UrduNepaliEnglishParallelCorpus.htm*

[5] *The list of the Penn-English Treebank files whose parallel Urdu translation is also available online can be found at: http://crulp.org/Downloads/ling_resources/parallelcorpus/Read_me_Urdu.txt and also at: http://crulp.org/Downloads/ling_resources/parallelcorpus/read_me_Extended_Urdu.txt only the files whose names are listed on these websites are used in this study.*

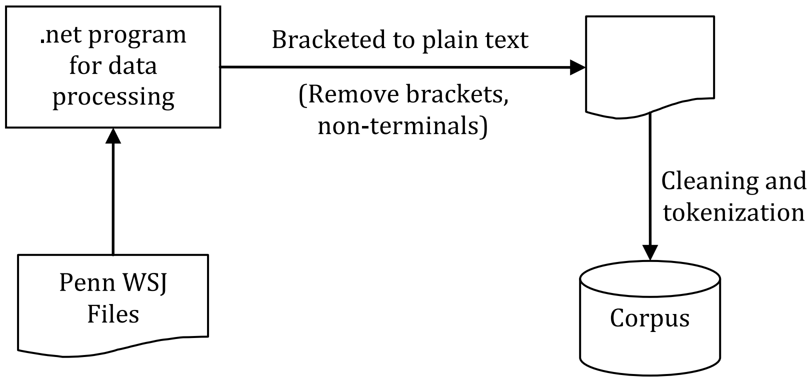


Figure 2.2: Overview of Corpus Creation from Penn Treebank-3 Corpus

As we have already mentioned above that Penn-Urdu corpus is available in plain text format and it's also sentence by tree aligned with the Penn-English Treebank, so we don't need to do any processing for the text extraction. By using the .net program, we convert the bracketed Penn-English Treebank into plain text data. This work is simply done by removing all the brackets from the data as well as non-terminal, the left over data is the terminal nodes of the tree that in order makes the sentence. Each WSJ file has multiple sentences in the tree format. To match the sentence format with Urdu Penn-Treebank data, we split the sentences over the part of speech tag ".", that marks the punctuation markers "." and "?". After getting plain text data cleaning and tokenization is performed, this is briefly discussed in Section 2.1. The summary of the whole process is as follows: our program strips off all the irrelevant tags and non-terminals, adds new line after processing each tree and at the end creates a plain sentence-aligned, text file. The plain-text file is then cleaned and tokenized, and whole process results in sentence aligned parallel corpus.

2.1.3 Quran and Bible Corpora

There are many online resources where Quran (Holy Book of Muslims) is easily available in both English and Urdu languages in UTF8 format, we selected an online resource[6] from several others where parallel data is freely available to download. The problem we encountered for downloading the Quran's data is data format; at most sites it is only available for downloading in image and xls format. For that reason we crawled the web link to get the plain UTF8 text data. We also found the

[6] *The Quran-English UTF-8 data is downloaded from: http://www.irfan-ul-quran.com/quran/english/contents/sura/cols/0/ar/0/ur/0/ra/0/en/1/ and, Quran-Urdu UTF-8 data is downloaded from: http://www.irfan-ul-quran.com/quran/english/contents/sura/cols/0/ar/0/ur/1/ra/0/en/0/*

free online resource of Bible (Holy Book of Christians). Bible's several versions in English are available but we could only get the parallel translation of the New Testament. Bible's English to Urdu data is not easily available, we hardly manage to find only single resource[7] where Bible's bilingual data is available in UTF8 format, otherwise Bible data is only available on the web in image and other non-UTF8 formats. After finding the resources of parallel corpora we extracted the bilingual corpus using the *self-written* java based Web Crawler[8].

Crawler's implementation is generic for getting both monolingual and bilingual data. So we have made the modifications in the crawler for extracting the parallel corpus. The generic implementation works this way: we provide the main website link to the crawler; it collects all the links from the main pages and adds them into its repository and also extracts the data from main page and stores it.

The links from the repository are fetched one by one and again the same process is repeated until all the sub-links are accessed exactly once. This generic implementation worked for the monolingual data collection as we just want to collect all the available Urdu data from the links. For the parallel corpus collection we first analyze the format of the links that contain the parallel data and we only add those links in the crawler repository that contains the parallel data and simply crawl the data from the stored links in the repository and don't add newly encountered links in the repository.

The Quranic data is available in the form of the *Suras*[9], each Sura consist of minimum 3 to maximum 286 sentences. There are 114 total Suras in the Quran, so all together we crawled 114 pages for each language to build Quran's bilingual corpora. Whereas Bible consist of 27 chapters where English data is dumped from one single html page and Urdu data is crawled from the 27 sub-links of the main link already provided above.

The data extraction procedure of Quran's bilingual data and Urdu version of the Bible is illustrated in Figure 2.3 and described below. The data extraction procedure in the pipeline above is quite similar to the process already described for Emille corpus creation. The only difference is in the first phase of the pipeline where in Emille the bilingual data was provided and here we crawled the data from the online resources. The process works as follows: we feed the main web link into the crawler and define the format of the dynamic creation of the rest of the sub-links where bilingual data is stored. Crawler builds the web link repository and starts fetching the data from the links one by one. Data gets cleaned in the next step, all the html tags are removed, blank lines are deleted and data gets

[7] *The free King James Bible edition is distributed by "Project Gutenberg Etext". The Bible-English UTF-8 data is downloaded from http://www.gutenberg.org/dirs/etext90/kjv10.txt And, the Bible-Urdu UTF-8 data is downloaded from: http://www.terakalam.com*

[8] *The web crawler was specially written for the corpus collection for this study work.*

[9] *T*he Chapter in the Quran is known as Surah.

stored on the disk in the 2nd phase. At the end of the 2nd phase we ended up with unaligned parallel corpora. In the next step the Bible's corpus is manually aligned sentence by sentence. After manual alignment data gets cleaned and tokenized to create final corpus. Data cleaning is already discussed in Section 2.1 and data tokenization is discussed in section 4.1.

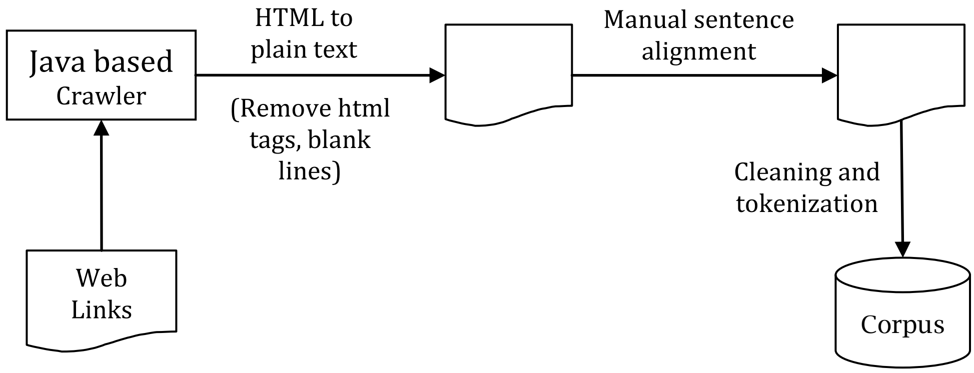


Figure 2.3: Overview of Quran and Bible Corpus Creation from the Web resources.

2.2 Collection of Monolingual Data

Large amount of Urdu data that consists of flat sentences is collected for the purpose of the study conducted for this research work. The monolingual corpus is used to make the language model that is used by the decoder to figure out which translation output is the most fluent among several possible translation options. Because of this fact language model of million tokens needs to be created to get better translation output. In this study we also tried to gather huge monolingual data from as many different available online sources as possible. The next step is to train the language model on the corpus that is suitable to the domain. To fulfill this need, data from diverse domains is collected. The main categories of the collected data are News, Religion, Blogs, Literature, Science, Education and numerous others. The lists of sources[10] used for data collection are as follows: BBC Urdu[11], Digital Urdu Library[12],

[10] *All the sources provide free E-Text with the requirement of proper mentioning the references of the material used and also the material can be used for the* ***non-profit making*** *research work.*

[11] *http://www.bbc.co.uk/urdu/*

[12] *http://www.urdulibrary.org/index.php?title=صفحہ_اول*

ifastnet[13], Minhaj Books[14], Faisaliat[15] and Noman's Diary[16]. The target side of the parallel corpora is also added to the monolingual data.

The data collection from the sources listed above and further processing on data was performed in three main steps that are described below:

i. Data Crawling and Processing: After collecting the list of available sources for free text we crawled the web-links using the crawler discussed in detail above. After getting the html pages we extract the data and remove all the html content from the text. We also remove all blank lines at this stage to limit the size of the data so that difficulty for processing the large amount of data can be avoided.

ii. Language Detection: The data extracted from the web was not completely in Urdu language, it contains languages other than Urdu and that makes data unusable. Mostly data included text in Arabic and English. To resolve this process we used the Perl script named LanguageDetector.pl[17], for detecting the languages other than Urdu and remove them from the data. Our Script doesn't delete the words from the middle of the sentences that will leave the data ungrammatical; rather it deletes the whole sentence if the proportion of the words belonging to the language other than Urdu is more than the words in Urdu.

iii. Cleaning and Tokenization: In the normalization step we removed bad characters and extra spaces from the data. Whereas tokenization process is the same as applied to the bilingual corpora.

2.3 Statistics over Corpora

This section provides the brief overview and description of the data used in this study. It summarizes the statistics over the raw corpora.

2.3.1 Parallel Corpora

The statistics over the bilingual corpora are summarized in Table 2.1 and Table 2.2. These corpora consist of plain sentences and they are

---

[13] *http://kitaben.ifastnet.com/*

[14] *http://www.minhajbooks.com/urdu/control/Txtformat/یونیکوڈ-کتب.html*

[15] *http://shahfaisal.wordpress.com/*

[16] *http://noumaan.sabza.org/*

[17] *LanguageDetector.pl is the corpus pre-processing utility script that statistically identifies the language of the given word based on the suffix. This script is written and kindly provided by Amir Kamran.*

constructed for the purpose of the study conducted for this research work. Corpora are used to induce phrase translation tables that are consulted by the decoder to figure out how to translate input in one language into output in another language. The part of these corpora is also used for parameter tuning and testing the translation output.

Corpus	Source	# of Sentences	# of Tokens	Vocabulary Size	Sentence Length	
					μ	σ
Emille	ELRA	8,736	153,519	9,087	17.57	9.87
Penn Treebank	LDC	6,215	161,294	13,826	25.95	12.46
Quran	Web	6,414	252,603	8,135	39.38	28.59
Bible	Gutenberg	7,957	210,597	5,969	26.47	9.77

Table 2.1: English Parallel Corpus Size Information

The size of the corpora approximately ranges from one hundred thousand to two hundred thousand tokens. The Emille corpus is the largest corpus in terms of the number of the sentences and it has the 2nd highest vocabulary size in all the corpora but it contains the least number of tokens among all the corpora. Penn Treebank has the highest vocabulary size. Intuitively we can consider the richness and rather fine-grain granularity of news domain. Conversely it is the smallest corpus among all the corpora in terms of the number of sentences. Bible has the 2nd maximum number of sentences in all domains but Bible and Quran have the minimum vocabulary size rate and that indicates the tendency of limited vocabulary usage in this domain. Corpora from the religious domain have the maximum number of tokens, followed by the Penn Treebank.

Corpus	Source	# of Sentences	# of Tokens	Vocabulary Size		Sentence Length	
				Raw Text	Normalize	μ	σ
Emille	ELRA	8,736	200,179	10,042	9,626	22.91	13.07
Penn Treebank	CRULP	6,215	185,690	12,883	12,457	29.88	14.44
Quran	Web	6,414	269,991	8,027	7,183	42.09	30.33
Bible	Web	7,957	203,927	8,995	6,980	25.62	9.36

Table 2.2: Urdu Parallel Corpus Size Information

The statistics of target side of bilingual corpora that is shown in Table 2.2 also concludes almost the same results for all corpora as drawn from the source side of the parallel corpora except the number of tokens in Urdu side of the Emille corpus are more than the numbers of tokens in Penn-Urdu Treebank. The most interesting phenomenon in comparison of both English and Urdu parallel corpora is that in all corpora except Bible, the number of tokens in Urdu corpora are more than the English corpora which is usual. But, in Bible the numbers of tokens in Urdu corpus are less than the number of tokens in English corpus. This could be because of difference in translation style since we are using different sources for both English and Urdu Bible corpuses. Another possibility is the different approach of language expressivity is adopted for Bible's Urdu corpus i.e. minimum usage of words to convey the meaning.

We also have summarized the change in vocabulary size after applying the normalization process. Emille and Penn have smaller loss in vocabulary size after applying the normalization, while Bible corpus has decrement of around 2000 unique words. This shows the wrong usage of diacritic marking and even there are chances of marking multiple entries of the same word differently. Examples of the same word with different forms (different diacritic marking or even without diacritic marking) from the un-normalized Bible corpora are shown in shown Example 2.1.

Example 2.1:

(a) The translation of word “who” without diacritic marking in bold.

English Sentence:	And **who** is he that will harm you, if ye be followers of that which is good?
Urdu Sentence:	اگر تُم نیکی کرنے میں سرگرم ہو تو تُم سے بدی کرنے والا **کون** ہے؟
Transliteration:	agar tum nīkī karne men sargaram ho to tum se badī karne wālā **kaun** he?

(b) The translation of word “who” with pesh (ُ◌) diacritic mark.

English Sentence:	Then said they unto him, **who** art thou?
Urdu Sentence:	اُنہوں نے اُس سے کہا تُو **کُون** ہے؟
Transliteration:	unhoñ ne us se kahā tū **kaun** he?

(c) The translation of word “who” with zabar (َ◌) diacritic mark.

English Sentence:	And **who** shall be able to stand?
Urdu Sentence:	اَب **کَون** ٹھہر سکتا ہے؟
Transliteration:	ab **kaun** ṭhahar saktā he?

In Example 2.1, Urdu variant of word “*who*” has three different possible forms and among those forms only forms in Example 2.1 (a) and (b) are correct. The real form of the word “*who*” is provided in Example 2.1 (b) whereas mostly Urdu literature is written and understandable without diacritic marking so because of that reason, word form in Example 2.1 (a) is also correct.

The vocabulary size of all normalized Urdu corpora is around 1000 words more than the vocabulary of English corpora except the source Penn Treebank corpus whose vocabulary size is around 1400 words more than the Urdu parallel corpora.

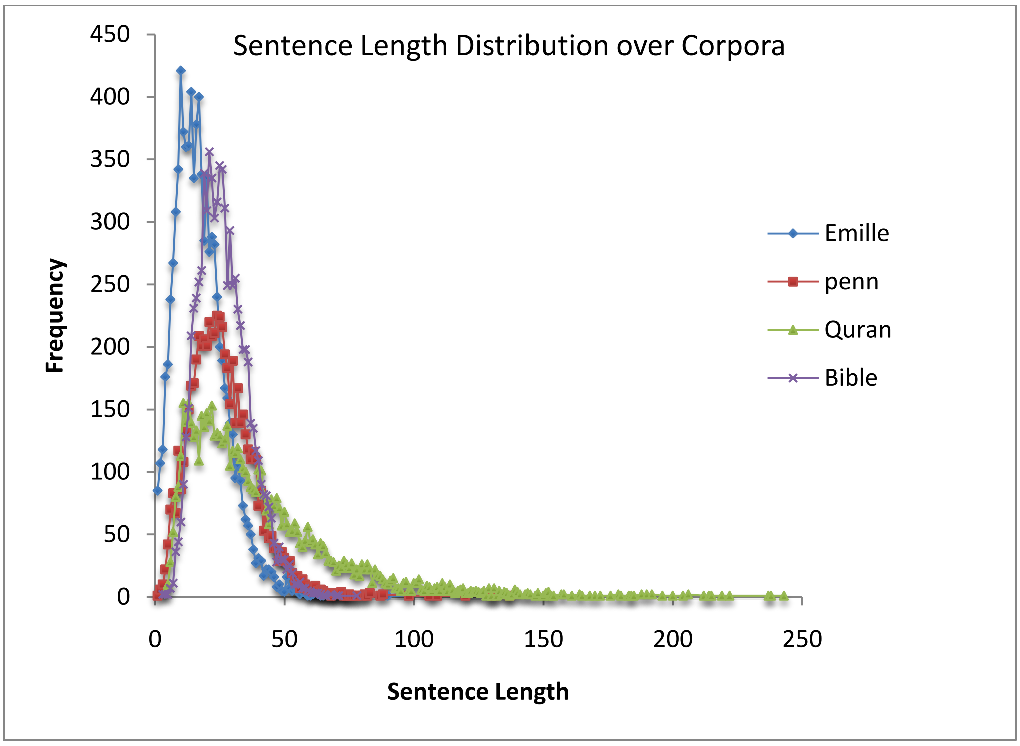


Figure 2.4: Sentence Length Distribution over the English side of bilingual Corpora

As for average sentence length, the average sentence length varies across the corpora. It is between 8 to 39 words on average for English side of parallel corpora and 23 to 42 words on average for Urdu side of the parallel corpora. The *Quran* corpus contains the longest sentence on average, while the *Emille* corpus has the shortest, whose average size is half of the sentences of the religious domain. The sentence length distribution over source side of bilingual corpus is illustrated in Figure 2.4 and for the target side of the corpora is illustrated in Figure 2.5.

In Figure 2.4 we can see that the average sentence length over all distribution is roughly around 25 words, and that the *Quran* corpus contains a few extraordinarily long sentences, with a size of even around 240 words. While, in Urdu corpora the sentence length over all distribution is roughly around 30 words and the maximum sentence length consists of around 260 to 270 words.

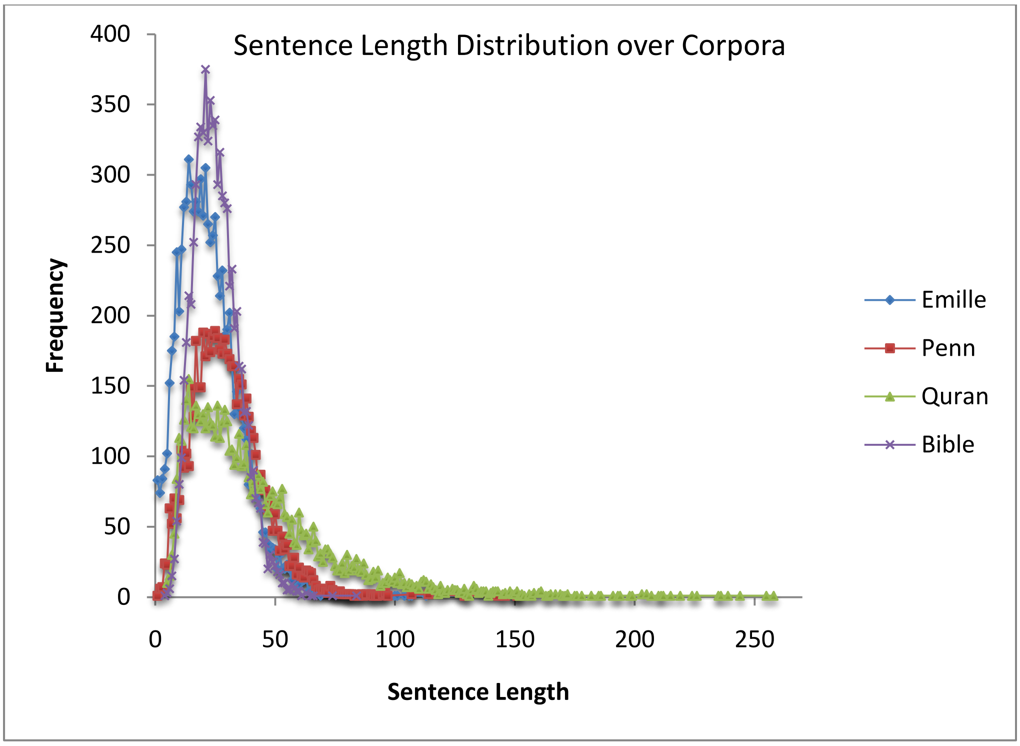


Figure 2.5: Sentence Length Distribution over the Urdu side of bilingual Corpora

2.3.2 Monolingual Corpora

The monolingual corpora collected for this study have around 61.6 million tokens distributed in around 2.5 millions sentences. These figures cumulatively present the statistics of all the domains whose data is used to build the language model. The language model for this study is trained on 62.4 million tokens in total and around 2.5 million sentences. This statistics is after adding in the monolingual data the target side of all the parallel corpora we collected for this study.

2.4 Data Normalization

The data we have collected in this study doesn't belong to any single organization and the various organizations that own the data have their own data formats or writing styles. For that reason, in all bilingual corpora, the Urdu corpora are written based on different writing standards. The main dissimilarities in writing style are as follows:

- Use of both English and Urdu punctuation markers.
- Diacritic marks usage in some of the corpora whereas rest doesn't prefer to use diacritics.
- Some corpora adopted to write numbers and dates in English numerals whereas some write them in Urdu numerals.

The un-normalized data would impact the translation system because of the obvious reasons; if in the system we have same word with different forms, translation system will treat them different words and this will lower down the probability of the correct Urdu translation against the English word.

For instance in case of diacritic marking same word with different form of the words as shown in Example 2.1, and for numerals same number is written half of the time in English format and sometimes in Urdu format. The list of English and Urdu numerals is provided in Table 2.3.

English Numeral	Urdu Numerals
0	٠
1	١
2	٢
3	٣
4	٤
5	٥
6	٦
7	٧
8	٨
9	٩

Table 2.3: Mapping between English and Urdu Numerals

In Table 2.4 we have shown the un-normalized sentence from Penn-Treebank corpus and also its modified form after applying the normalization steps.

Un-Normalized Urdu Sentence	۱۹۹۷ تک کینسر کا سبب بننے والے ایسبسٹاس کے تقریباً تمام با قیماندہ استعمالات کو غیرقانونی قراردیا جائے گا۔
Normalized Urdu Sentence	1997 تک کینسر کا سبب بننے والے ایسبسٹاس کے تقریبا تمام باقیماندہ استعمالات کو غیرقانونی قراردیا جائے گا۔
Transliteration	1997 tak kīnsar kā sabab banane wāle īsbasṭās ke taqrībā tamām bāqemāndah astaʾamālāt ko ğīrqānūnī qarār diyā jāe gā.

Table 2.4: Urdu Sentence from Penn corpus before and after applying normalization

To check the impact of different writing styles we performed two types of baseline experiments, one with the raw text and one after applying normalization on the Urdu data. The detail of the experiments performed and the results and comparison are discussed in Chapter 4.

2.5 Issues in Corpus

This Research study is very much dependent on the size and quality of parallel corpus. Unfortunately, when we started this work we couldn't find free English-to-Urdu, ready-to-use parallel corpus. That problem led us to create a parallel corpus by ourselves using all the available resources. After searching for all the data sources and writing the utilities to get plain text data out of the marked up data, we encountered the issues in the quality of the data as well as in the sentence-level alignment. In this section we describe those issues and the solutions of handling those issues.

Emille

Due to the multidimensionality of this corpus we decided to use the entire corpus for this study. But, we faced lots of issues in using its data. Not only there was problem in the data alignment but also the translation quality was very bad.

- As described above Emille data files contain multiple paragraphs and each paragraph contains multiple sentences. On analyzing the corpus we found that number of sentences in each paragraph is not same on both sides of the corpus because there were several sentences in the corpora without any translation at all.

- In some cases, the numbers of the lines in the paragraph on both sides of corpora were the same but the parallel sentence doesn't correspond to each other. We tried to deal with this issue and the problem explained above by aligning the both sides of the corpora.

- Among the numerals used in the entire corpus, 90% of the numbers (that are used as reference to the pages in manual) were not the correct match of each other in source and target side of the corpus. This issue could indeed cause the translation system to always output the wrong translation of numbers during testing. To remove this ambiguity we manually corrected all the numbers used in the corpora, so that each number in the source matches exactly the same number on the target side.

- In numerals mismatching, we also came across sentences that have numeral mismatch for numbers (other than reference to manual pages) are shown in Example 2.2.

Example 2.2:

English Sentence:	Have you been getting one of the following because of your illness or disability, in the last **26 weeks**?
Urdu Translation:	آپ کو پچھلے **182 دنوں** م یں اپنی بیماری یا معزوری کے سبب مندرجہ ذیل میں سے کوئی ایک ملتا رہا ہے ؟
Transliteration:	āp ko pičhle **182 dinoñ** meñ apnī bīmārī yā maʾazūrī ke sabab mandarjah żīl meñ se koī ek miltā rahā he?

In this example, 26 weeks is translated as 182 days in parallel corpus, problem words are shown in bold face.

The Urdu corpus of Emille also contains words from Sanskrit[18] vocabulary. A few of those words are not part of the Urdu vocabulary and not known by the native Urdu speakers. We also tried to replace the Sanskrit words with their Urdu equivalents. Some of the Sanskrit words that are changed in the corpus are provided in Table 2.5.

Gloss	weipārī	jānkārī	soč wičār
Sanskrit Word	ویپاری	جانکاری	سوچ وچار
Gloss	tājir	mʾalomāt	soč bičār
Converted Urdu Word	تاجر	معلومات	سوچ بچار

Table 2.5: Sanskrit expressions in Emille Corpus mapped on Urdu Vocabulary

We also found spelling mistakes in Urdu side of the parallel corpora. They are two different trends for the spelling mistakes found in the corpora. Firstly, wrong spelling is used throughout the corpus and secondly, the spelling is wrong in half of the corpora and half of the time its correct form is used.

[18] *It is a historical Indo-Aryan language and it is one of the 22 scheduled languages of the India.* ***Typical*** *Sanskrit vocabulary is not used in spoken and written Urdu Language.*

Correct Spelling	Wrong Spelling
بچے	بیچے
بچت	بیچت
الگ	لگ
تاہم	تا ہم

Table 2.6: Spelling mistakes in Emille corpus

In Table 2.6, the fourth word has spelling error due to the use of extra space between both constituents of the word.

Emille data is already very small and due to the lack of data we didn't feel it feasible to run any automatic alignment tool, because alignment tools not only delete the unaligned data but also aligned output is not very reliable. Due to the issues discussed above we decided to manually align the whole corpus and the output result of this process is *manually aligned whole Emille corpus*. In this available short time we also tried to improve the translation quality so around 25-30% of the sentences are also manually corrected (by making modifications in the sentence or rewriting the whole sentence). Most of the modifications are made on the English side of the parallel corpus.

Quran and Bible

Although parallel religious data is mostly sentence by sentence aligned but after data extraction and processing, because of some unknown reasons, we found some misalignments in the data. Due to only 2 to 3 unaligned sentences we had to manually analyze the entire corpora and find the proper locations in the corpora with mismatch sentences. Output of this phase is the sentence by sentence aligned corpora ready for the cleaning process.

Summary

In this chapter we presented the English-Urdu parallel and monolingual corpora collection in detail. We further explained the procedure of extracting the actual parallel text out of collected corpora and provided statistics of both parallel and monolingual corpora. We also presented the need of normalizing the target Urdu corpora and also the issues faced during and after the corpus collection. In following chapter we present our translation improvement techniques for the selected language pair.

Chapter 3

Improvement Techniques

This chapter starts with the discussion of the possible translation issues within the domain of phrase-based machine translation systems between the source and target languages selected for this study. Then, we present our target approach for the improvement in the quality of the translation obtained using phrase-based MT. We also explain the improvement techniques and the necessary tools required to apply those techniques. We further discuss the exploitation of some advanced features of the phrase-based system for dealing with the data sparseness problem occurs due to presence of highly inflected languages.

3.1 Selection of Translation Model

Before selecting the translation model for our study we discuss the few requirements to produce the translation for the selected language pair. The TM should provide the efficient word reordering model as English and Urdu have different word ordering structures and also it must be able to deal with the data sparseness problem, as Urdu is highly inflectional language and we never have a huge amount of data available that covers all possible forms of single word.

For this research study, after analyzing the requirements of selected language pair we decided to use the MT system based on phrase-based translation model, where phrases consist of words only. The major reason of selecting phrase-based MT is due to the faster training method and less computationally expensive model (within the domain of limited word reordering) as compared to other syntax-based MT systems. "More sophisticated approaches that make use of syntax do not lead to better performance. In fact, imposing syntactic restrictions on phrases, as used in recently proposed syntax-based translation models (Yamada, et al., 2001), proves to be harmful." (Koehn, et al., 2003) Syntax-based MT systems are slow to train and decode because the syntactic annotations further add a level of complexity.

For this study we preferred to use state-of-the-art phrase-based MT over hierarchical phrase-based MT due to the fast speed and reasonable memory requirement. Although hierarchical PBT system provides the syntactic reordering over the phrases but they are not very good at long-distance reordering. We try to utilize the fast and simple phrase-based architecture together with the reordering approach of syntax-based MT systems by preprocessing the source data. (Bojar, et al., 2008) and (Ramanathan, et al., 2008) used a similar technique for the English-Hindi language pair that is structurally similar to English-Urdu. Both have achieved a significant improvement after applying the preprocessing on source corpora. Another reason of selecting state-of-the-art phrase-based MT systems is the further extension of phrase-based translation models into factored based translation model (Koehn, et al., 2007) that helps in dealing with data sparseness issue and also helps in getting the grammatically coherent translation output.

3.2 Techniques

After considering the possible translation issues with the selected language pair and selecting the translation model according to those translation issues, we finally propose our techniques for improvement in translation. In this study we are using two different improvement techniques: dealing with the difference in word order of source and target languages and also attempting to deal with the issues due to richer morphology of the target language. The first technique applies the word order transformation over source language structure by preprocessing the data and second technique uses the factorized translation model for the translation.

3.2.1 Reordering

As explained in section 1.1, English is SVO language and Urdu follows (mostly) SOV structure. For translation from English to Urdu we need SMT system to perform long distance reordering to get the better translation output. Phrase-based systems can perform long-distance reordering using distortion models but, allowing long-distance reordering explodes the search space (i.e. too many possible partial hypothesis) beyond reasonable stack limits. So, the system has to decide prematurely and it is likely to lose good partial hypotheses in initial searching, hence causes the much higher risk of search errors.

To overcome this problem we preprocess the English training, development and test corpora prior to the SMT training and decoding cycle, and try to minimize the difference in word order of both languages using our scheme.

Transformation System

We have used the subcomponent of rule-based English-to-Urdu Machine Translation system (RBMT) (Ata, et al., 2007) for the preprocessing of only English corpus in the parallel corpora. We developed this RBMT system as the final project under Bachelor studies. In the different analysis levels of MT systems, our RBMT system falls under the transfer approach[1].

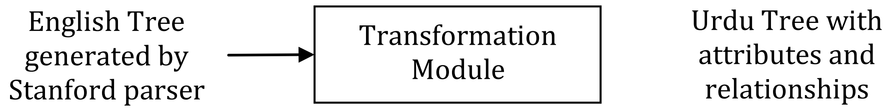


Figure 3.1: Transformation Module in Rule base English to Urdu MT Engine

There are three main components of this MT Engine: Dictionary, Transformation and Translation. For this study we use the Transformation module of the MT engine, shown in Figure 3.1 for transforming the structure of the English sentence according to the word order of the Urdu language. Our MT engine uses the open source API of the Stanford Parser[2] to generate parse tree of the English sentence. The generated parse tree is later passed as the input to the transformation module. This module uses the transformation rules and transforms the English tree into its equivalent Urdu-like tree. The transformation rules are kept separated from the transformation module so that the module can easily be adapted for any other language from the same family as Urdu that has the same structure but differs only in the transformation scheme. The rules can be easily added and deleted through an XML file.

The output tree from this module is not actually an Urdu tree; it's only a transformed English tree into Urdu sentence structure. For this study, we have modified the transformation module according to our needs. We pass the English parse tree and apply the transformation rules on the parse tree to get transformed English tree, none of the attributes and relationships are retrieved during this process.

Stanford Parser

The Stanford parser takes the English sentence, parses the sentence using a Probabilistic Context free grammar and outputs the parsed tree.

---

[1] *In the transfer approach, the translation process is decomposed into three steps: analysis, transfer and generation. In the analysis step, input sentence is analyzed using parsers and/or morphological tools, producing abstract representation of source sentence. In the transfer step, this representation is transferred into the corresponding representation in the target language. In the generation step, the target language sentence is generated.*

[2] *http://nlp.stanford.edu/software/lex-parser.shtml Stanford parser is also available online at: http://nlp.stanford.edu:8080/parser/*

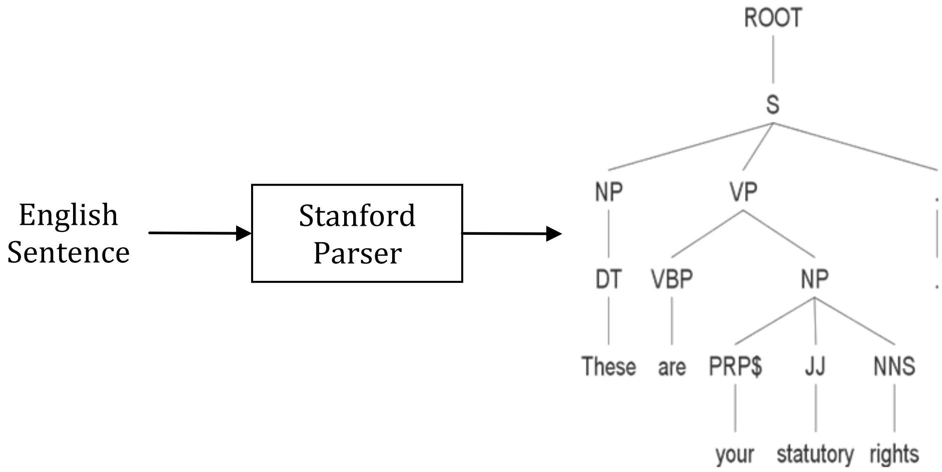


Figure 3.2: Stanford Parser API's input and output format

Transformation Rules

Transformation rules are the key element of our transformation system. The rules defined for the RBMT system are based on reverse Paninian grammar theory. In this system mapping rules are defined by just reversing the order of the constituents of the linguistic phrases (NP, VP, etc) with a few exceptions. Example 3.1 shows, if the grammar rule in English sentence for verb phrase consists of a verb phrase and an object NP then its corresponding Urdu transformation rule consists of reverse ordering of constituent phrases in grammar rule. The (*) in grammar rules is used for the purpose of generalization. For instance, according to the form of the verb Stanford parser uses different tag sets for representing verb node; VB for the verb basic form, VBZ is used for basic verb form with third person singular and many others. To cover all possible tags for verb node in each grammar rule, we use the generalized grammar rule instead of writing same grammar rule with every possible POS tag. The output of RBMT system is grammatically coherent Urdu translation.

Example 3.1:

Grammar Rule: $VP \rightarrow VB^* NP$[3]

Transformation Rule: $VP \rightarrow NP\ VB^*$

In Example 3.1, VP corresponds to the Verb Phrase, VB represents the Verb node and NP matches with the object Noun Phrase node.

---

[3] *For further detail about the POS tags, refer to the Stanford POS Tag set.*

The rule defined in Example 3.1 can be added into the system without actually writing the complete transformation rule. Example 3.2 shows how rules are basically added into the system for the sake of simplicity.

Example 3.2:

```
<rule>
    <english-rule>VP -> VB* NP</english-rule>
    <urdu-transformation>reverse</urdu-transformation>
</rule>
```

Example 3.1 and Example 3.2 shows exactly same grammar rule and its corresponding transformation rule. The only difference is the format of the transformation rule. If a transformation rule is formed by exactly reversing the ordering of the constituent nodes then, the transformation rule is defined by just writing the string "reverse" instead of writing complete transformation rule in reverse order. If a grammar rule consists of more than 2 constituent nodes and the transformation rule doesn't correspond to the exactly reverse ordering of constituent nodes in the grammar rule then, Example 3.3 shows the design of the transformation rule for representing the ordering of the constituent nodes in the transformation rule by actually marking them with their indexes in the grammar rule.

Example 3.3:

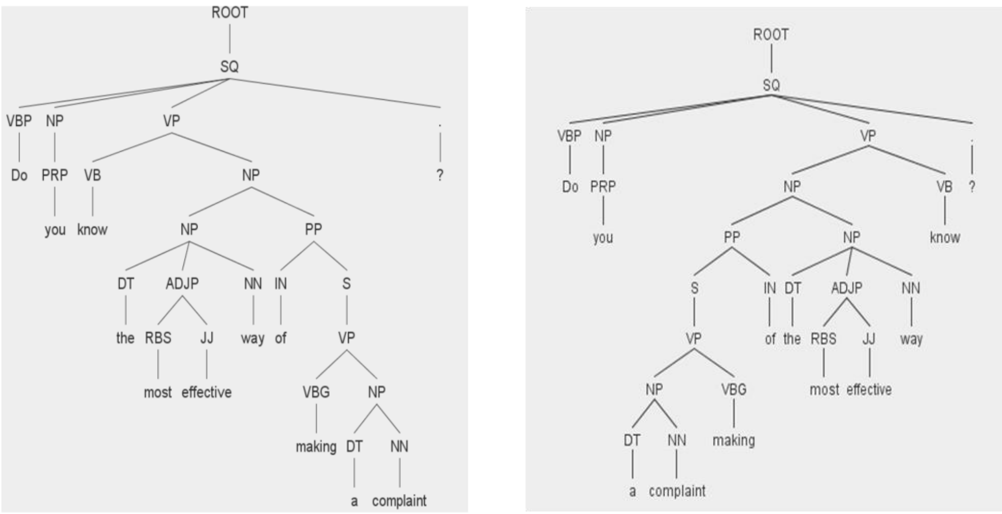
```
<rule>
    <english-rule> VP -> VB ADVP PP </english-rule>
    <urdu-transformation>1 2 0</urdu-transformation>
</rule>
```

In Example 3.3, the grammar rule with constituents VB, ADVP and PP corresponds to the order 0, 1 and 2 respectively. The transformation rule with numbered ordering represents the rule VP $\rightarrow$ ADVP PP VB. The default ordering of rules also needs to be defined. For instance, if transformation rule does not exists then the default rule for that grammatical category will be used. Example 3.4 shows the default rule for VP.

Example 3.4:

```
<rule>
    <english-rule> VP -> default</english-rule>
    <urdu-transformation>reverse</urdu-transformation>
</rule>
```

Figure 3.3 shows an English parse tree with its transformed Urdu tree using the transformation rules.



English Parse Tree Transformed Tree

Figure 3.3: English Parse Tree from Stanford Parser with Transformed Tree

Most of the transformation rules are formed by reversing the order of the constituents in grammar rule. But, there are a few exceptions in which order is not reversed and transformation rules more or less follow the ordering of the grammatical rules.

- Adjectives are followed by nouns (if exist).
- In question sentences, question word comes at the beginning of the sentence.
- Adjectives are preceded by adverbs (if exist).
- Adverbs are placed before verbs (mostly).

Extension in Transformation Rules

As transformation rules in RBMT system are generated by following the theoretical model of reverse Panini grammar so, for capturing the most commonly followed word order structures in Urdu language we defined a new set of transformation rules required for word order transformation. For this study we perform analysis on parallel corpora and accumulate the transformation rules representing the most frequent ordering of constituents in phrase structures.

For this study we gather a set of around 90 to 100 transformation rules.

In Example 3.5 we are showing some transformation rules that are analyzed and created for this study.

Example 3.5:

Prepositions become postpositions

Grammar Rule: $PP \rightarrow IN\ NP$

Transformation Rule: $PP \rightarrow NP\ IN$

Verbs come at the end of sentence and ADVP are followed by verbs.

Grammar Rule: $S \rightarrow ADVP\ VP\ NP$

Transformation Rule: $S \rightarrow NP\ ADVP\ VP$

The effect of preprocessing the English corpus and its comparison with the distance reordering model are discussed in Chapter 4 in detail.

3.2.2 Factorization

In SMT systems each form of the word is treated as independent entity, this problem gives rise to the data sparseness issue that is caused by limited training data. Due to data sparseness, languages having rich morphology negatively influence the MT performance. With the use of morphological information, the requirement for the large training data can be reduced. Recent phrase-based MT systems are now further extended to factor-based models that interpret each entity as a factor instead of single token (word). Factor in the MT systems represent the vector of different level of annotations added at the word level. For example in factored model each factor can consist of word, lemma, part-of-speech, morphology, etc.

Due to the limited availability of resources for Urdu, we are unable to integrate morphology in the system. Instead, our factored model will operate on word, lemma and pat-of-speech. We also use the additional n-gram language model over the POS tags. To start with the real experiments on English-Urdu factored model we require the linguistic tools i.e. lemmatizer and POS tagger for both English and Urdu.

Tools for English

For this study we use the Stanford lemmatizer and Stanford Maximum Entropy Part-of-Speech tagger[4] (Toutanova, et al., 2000) for annotating English corpora. The Stanford tagger uses the Penn Treebank tagset for POS tagging. In this study, we are using the already trained bidirectional-

[4] *http://nlp.stanford.edu/software/tagger.shtml*

distsim-wsj-0-18.tagger model (provided with the tagger) for tagging English data. This model is trained on WSJ sections 0-18 using bidirectional architecture, including word shape and distributional similarity features. The trained tagger model has accuracy 97.28% correct on WSJ 19-21 (90.46% correct on unknown words).

Lemmatizer is provided in the tagger package inside the process. Morphology directive. Table 3.1 shows the input provided to the Stanford tagger and the output generated by the tagger. Output is represented as "word | lemma | POS-tag".

Input	Do you know the most effective way of making a complaint?
Output	Do|do|VBP you|you|PRP know|know|VBP the|the|DT most|most|RBS effective|effective|JJ way|way|NN of|of|IN making|make|VBG a|a|DT complaint|complaint|NN ?|?|.

Table 3.1: Stanford tagger's Input and Output for factored Model

Tools for Urdu

Very little effort has been put for the development of linguistic tools for Urdu language analysis. The tools specifically dedicated to the Urdu language analysis are developed by the research institute, Centre for Research in Urdu Language Processing[5] (CRULP). There are a few drawbacks associated with the tools provided by the CRULP.

- Complete Documentation is not provided with the tools.
- (Often) the input and output format of the tools make it hard to use.
- Accuracy of the tools (except POS tagger) is not mentioned in the (limited) documentation.
- Statistical tools cannot be retrained.

i. POS Tagger

Because of above mentioned reasons, we first decided to train a model of Stanford tagger for Urdu data provided by CRULP using the manually tagged WSJ 00-02, 317 stories from start. For training the Stanford tagger properties file is required with the few essential parameters. For example model, trainFile, arch, etc. The statistics of data used for training and testing Stanford tagger for Urdu is shown in Table 3.2.

[5] *http://www.crulp.org/*

Tagged Training Data		Testing Data	
Total Sentences	Total Words	Total Sentences	Total Words
5822	167673	457	12156

Table 3.2: Statistics of Penn Treebank data used for training and testing Stanford Tagger for Urdu

The Stanford tagger was tested on 12156 words from which, 4285 were found to be unknown. The accuracy of the trained tagger is in detail provided in
Table 3.3.

	Correctly Tagged	Wrongly Tagged	Total Count	Accuracy (%)
No. of Sentences	2	455	457	0.437
No. of Words	3005	9151	12156	24.72
Unknown Words	0	4285	4285	0

Table 3.3: Accuracy of Stanford trained model for Urdu Tagger

Table 3.3 shows the input sentence and the tagged output sentence generated by Stanford's trained tagger. The reference tagged sentence is also provided.

Input	بیل ، لاس اینجلیس میں واقع ، برقیات ، کمپیوٹر اورتعمیری مصنوعات بناتا اورتقسیم کرتا ہے ۔
Transliteration	bīl, lās injlīs meñ wāqaʾa, barqyāt, kampyūṭar aur taʾamīrī mṣnūʾāt banātā aur taqsīm kartā he.
Reference	بیل/NNP ،/PM لاس/NNP اینجلیس/NNPC میں/CM واقع/JJ ،/PM برقیات/NN ،/PM کمپیوٹر/NN اور/CC تعمیری/JJ مصنوعات/NN بناتا/VB اور/CC تقسیم/NN کرتا/VB ہے/AUXT ۔/SM
Output	بیل/JJ لاس/JJ اینجلیس/JJ میں/CM واقع/JJ برقیات/JJ کمپیوٹر/JJ اور/JJ تعمیری/JJ مصنوعات/JJ بناتا/VB اور/JJ تقسیم/JJ کرتا/JJ ہے/AUXT

Table 3.4: Reference, Input and tagged output sentence using Stanford Tagger

Due to only 24.7% accuracy of the Stanford tagger on test set, we decided to use CRULP's statistical POS tagger[6] with the 97.2% accuracy mentioned on their web-link. We already mentioned before a few problems in using the CRULP's tools. Another major issue associated with the CRULP's POS tagger is that the tagset used for building the training model is different from the tagset used for manually tagging the WSJ Urdu data. Consequently the accuracy of the tagger cannot be measured automatically, as tagsets of tagger and manually tagged data are different. On manually analyzing the tagged data produced by CRULP's tagger, results were not found to be satisfactory.

The example sentence in Table 3.5 is used from manually tagged corpus provided by CRULP and we assume that the same data should have been used for training a tagger for Urdu. Although POS tags in output and reference sentences don't have one-to-one correspondence as tagset is different but still we can see the difference in POS tags classes in both reference and output sentences.

Input	اکسٹھ سالہ پیئرونکن ۲۹ نومبر کو بطور نان ایگزیکٹو ڈائریکٹر بورڈ میں شامل ہوں گے۔
Transliteration	aksṭah sālah peironkn 29 nawambar ko baṭor nānāygzekṭū ḍāirekṭar borḍ meñ šāmal hūñ gae.
Reference	<CD>اکسٹھ <JJ>سالہ <NNP>پیئرونکن <CD>۲۹ <NNP>نومبر <CM>کو <RB>بطور <JJ>نان ایگزیکٹو <NN>ڈائریکٹر <NN>بورڈ <CM>میں <JJ>شامل <VBL>ہوں <AUXT>گے <SM>-
Output	<ADJ>اکسٹھ <NN>سالہ <VB>پیئرونکن <AA>۲۹ <PN>نومبر <P>کو <ADV>بطور <ADJ>نان ایگزیکٹو <NN>ڈائریکٹر <NN>بورڈ <P>میں <NN>شامل <VB>ہوں <AA>گے

Table 3.5: Output generated using CRULP's tagger

The difference in tagset can be seen clearly like JJ-ADJ, RB-ADV, PN-NNP and AA represents aspectual auxiliary. But on mapping the tags we will analyze that most of the content words are tagged incorrectly. Although the accuracy is claimed to be 97% but the results are not adequate to be used in this study. Accordingly, for this study we use the Statistical POS tagger[7] based on word suffixes with the accuracy approximately around 78.9% on test set. To increase the accuracy of the tagger we further added the close class words and cardinals and trained the tagger again on the same amount of the data presented in Table 3.2.

---

[6] *http://www.crulp.org/software/langproc/POS_tagger.htm*

[7] *POS tagger is written and kindly provided by Amir Kamran.*

<table>
<tr><td>Input</td><td>بیل ، لاس اینجلیس میں واقع ، برقیات ، کمپیوٹر اور تعمیری مصنوعات
بناتا اور تقسیم کرتا ہے ۔</td></tr>
<tr><td>Transliteration</td><td>bīl, lās injlīs meñ wāqaʾa, barqyāt, kampyūṭar aur taʾamīrī
mşnūʾāt banātā aur taqsīm kartā he.</td></tr>
<tr><td>Reference</td><td>بیل/NNP ،/PM لاس/NNP اینجلیس/NNPC میں/CM واقع/JJ ،/PM
برقیات /NN ،/PM کمپیوٹر/NN اور/CC تعمیری/JJ مصنوعات/NN
بناتا/VB اور/CC تقسیم/NN کرتا/VB ہے/AUXT ۔/SM</td></tr>
<tr><td>Output</td><td>بیل/NNP ،/PM لاس/NNP اینجلیس/NNP میں/CM واقع/JJ ،/PM
برقیات/NN کمپیوٹر/NN اور/CC تعمیری/JJ مصنوعات/NN بناتا/VB
اور/CC تقسیم/NN کرتا/VBL ہے/AUXT ۔/SM</td></tr>
</table>

Table 3.6: Output generated using Kamran's Tagger

The accuracy of the tagger increased from 78.9% to 79.4% on the same testing data used for Stanford's tagger testing.

ii. Stemmer

For factored translation, we use the stem form of each Urdu word to overcome the data sparseness. For this purpose we use the Urdu

<table>
<tr><td>Input</td><td>کیا آپ کے قانونی اختیارات کے بارے میں آپ کو صحیح معلومات
ہے ؟</td></tr>
<tr><td>Transliteration</td><td>kyā āp ke qānūnī axtyārāt ke bāre meñ āp ko şaḥeḥ
maʾalūmāt he?</td></tr>
<tr><td>Output</td><td>کیا|کیا آپ|آپ کے|کے قانونی|قانون اختیارات|اختیار کے|کے
بارے|بارے میں|میں آپ|آپ کو|کو صحیح|صحیح معلومات|معلوم
ہے|ہے ؟|؟</td></tr>
<tr><td>Transliteration</td><td>kyā|kyā āp|āp ke|ke qānūnī|qānūn axtyārāt|axtyār ke|ke
bāre|bāre meñ|meñ āp|āp ko|ko şaḥeḥ|şaḥeḥ
maʾalūmāt|maʾalūm he|he ?|?</td></tr>
</table>

Table 3.7: Stemmed Output using CRULP's Stemmer

Stemmer[8] provided by the CRULP. Fortunately for Urdu stemmer, CRULP has provided the Stemmer DLL that we use to run the stemmer on our Urdu corpus. The text input to the CRULP's stemmer and stemmed output is shown in Table 3.7. The output format is word | stem, from right to left. We then combine the tagged and stemmed output and formalize the data in the format that can be used with the factored model. In Table 3.8 we have presented the sample sentence from Emille corpus, where each token is represented as factor. The format of factor includes word | stem | POS tag.

Input	کیاآپ کے قانونی اختیارات کے بارے میں آپ کو صحیح معلومات ہے ؟
Output	اختیارات|اختیار|JJقانونی|قانون|KER کے|کے|PRRF آپ|آپ|QW کیا|کیا آپ|آپ |PR میں|میں| NN بارے|بارے | KER کے|کے |NN ہے|ہے |NN معلومات|معلوم|JJ صحیح|صحیح|CM کو|کو|PRRF AUXT|؟|؟|SM

Table 3.8: Factor format used for factor-based translation

Summary

In this chapter, we first introduced the translation model that is used throughout this study. Next, we discussed the issues in the selected translation model and presented the improvement techniques to overcome those issues. We also introduced the tools that are used for the improvement techniques, both for English and Urdu languages. This chapter concludes by looking at the specific improvement techniques, applicable to a phrase-based machine translation system, to improve the translation quality. In the following chapter we will present the experimental results after applying the techniques discussed in this chapter.

[8] *http://crulp.org/software/langproc/UrduStemmer.htm*

Chapter

4

Experiments and Results

This chapter presents the results of different set of experiments carried out for this study. The necessary detail of corpora that are used during the experiments is presented in Chapter 2. The chapter starts with the experimental setup, followed by the description of the evaluation measure used to evaluate translation output. The main part of the chapter focuses on presenting and discussing the improvements in translation quality, formally discussed in Chapter 3. The four major experiments conducted for this study are: baseline experiments, experiments with distance-based reordering, experiments after applying word order transformation and experiments using factored based model. The chapter concludes by comparing the results and translation quality of the generated output using different experimental setups.

4.1 Experimental Setup

In this section we describe the toolkit used for building the language model. We also illustrate about the translation system used for conducting the experiments. We further discuss the translation procedure, together with the different parameter settings adopted for carrying out the experiments. We also discuss in detail the data preparation for the different experiments.

4.1.1 Tools

In this section we provide the detail of the translation system used to perform translation between English and Urdu and also the necessary toolkits required together with the translation system.

The Statistical Language Modeling Toolkit

There are various software packages available to build Statistical Language Model. For example, the SRI Language Modeling toolkit[1]

[1] *http://www.speech.sri.com/projects/srilm/*

(SRILM) (Stolcke, 2002), or IRST Language Modeling toolkit[2] (IRSTLM) (Federico, et al., 2008)

In this study, we use **SRILM** (Stolcke, 2002). SRILM toolkit is composed of set of tools for building and applying Statistical Language Models (LMs). The main purpose of SRILM is to support Language Model estimation and evaluation. Estimation means the creation of a model from training data; evaluation means computing the probability of a test corpus (Stolcke, 2002).

For this study, we use the SRILM tool *ngram-count* to estimate two language models. One language model is built upon a text monolingual Urdu data by using Chen and Goodman's modified Kneser-Ney smoothing (Chen, et al., 1999). Second language model is comprised of part-of-speech tagged monolingual data, built using Witten-Bell discounting. We first tried to build the POS language model using Kneser-Ney smoothing technique but came across with smoothing issues[3], as KN-discounting is based on counts-of-counts i.e. number of words occurring once, twice, etc and in POS LMs the lower order n-gram counts are much fewer because there could be very few POS tags that occurs once or twice in a given corpus. For that reason different smoothing technique is used for building POS LM. The POS tagged LM is used together with text based language model in factor base translation model.

By default SRILM removes the unknown words in calculating the ngram-counts; we build the *open vocabulary* LM i.e. one that contains the unknown-word tokens as a regular word. SRILM can induce a language model of any order; in this study we have chosen to use the trigram language model unless stated otherwise.

Translation System

The statistical phrase-based machine translation system, **Moses[4]** (Koehn, et al., 2007), is used in this work to produce English-to-Urdu translation. According to (Koehn, et al., 2007) "The toolkit is a complete out-of-the-box translation system for academic research. It consists of all the components needed to preprocess data, train the language models and the translation models. It also contains tools for tuning these models using minimum error rate training (MERT) (Och, 2003)".

Moses automatically trains the translation models on the parallel corpora of the given language pair. It uses an efficient algorithm to find the maximum probability translation among the exponential number of candidate choices. For this study we have chosen to build phrase

---

[2] *http://hlt.fbk.eu/en/irstlm*

[3] *Issues in building POS tagged LM is discussed under SRILM FAQ section, Smoothing Issues. http://www-speech.sri.com/projects/srilm/manpages/srilm-faq.7.html*

[4] *http://www.statmt.org/moses/*

translation table on 7-gram of the words for each phrase, unless stated otherwise.

4.1.2 Translation Setup

The training process in Moses takes nine steps and all of them are executed using the script *train-factored-phrase-model.perl*. The training steps, external tools used for the training by Moses and also the parameters settings at each step are described below:

i. *Prepare Data: the selected corpus for the experiment is first cleaned using tok-dan.perl*[5] *script. It tokenizes the data and removes the redundant space characters. It also removes the extra spaces on the start and end of the line. The data is then converted to lowercase using the lowercase.perl script provided with the Moses implementation.*

ii. *Word Alignment: Moses uses GIZA++*[6] *(Och, et al., 2000) toolkit which is freely available implementation of IBM models for extracting word alignments. Alignments are obtained by running the toolkit in both translation directions and then symmetrising the two alignments. In our study we have used the grow-diag-final-and*[7] *alignment heuristic. It starts with the intersection of the two alignments and then adds additional alignment points that lie in the union of the two alignments. This method only adds alignment points between two unaligned words.*

iii. *Extract Phrase: Using the generated word alignment, Moses estimates the Maximum likelihood lexical translation table and extracts all those phrases in which words are aligned only to each other and not to any word outside the phrase.*

iv. *Score Phrases: Phrases are scored from the stored phrase translation table. For each pair five different phrase translation scores are computed:*

- *Phrase translation probability* $\emptyset\ (f|e)$
- *Lexical weighting* $lex(f|e)$
- *Phrase translation probability* $\emptyset\ (e|f)$
- *Lexical weighting* $lex(e|f)$
- *Phrase penalty (always exp(1) = 2.718)*

---

[5] *Tok-dan.perl is a low-level data tokenizer, written and kindly provided by Daniel Zeman*

[6] *http://www.isi.edu/och/GIZA++.html*

[7] *grow-diag-final-and works via expanding the alignment by adding directly neighboring alignment points and alignment points in the diagonal neighborhood.*

v. *Reordering: Moses builds the lexicalized reordering model that conditions the reordering on the actual phrases. It provides three different reordering models (i.e. different types of orientation of the phrases) together with number of variations of the lexicalized reordering model based on the orientation types. We have used in our experiments distance and msd-bidirectional-fe[8] reordering models. By default Orientation-bidirectional reordering model is used in all the experiments for building the reordering table. Along with bi-directional model, if the distance-based models are used then it is mentioned explicitly.*

vi. *End of Training: after creating reordering table, generation table is built using the target side of the training corpus. We have used different parameters for the building the generation table in factored based translation, for experiments with word reordering (only) default settings are used. Training ends with the successful creation of the configuration file called Moses.ini.*

After training the translation model, Moses standard MERT is executed on development set for tuning the weights of the individual models in our setup.

4.1.3 Data Preparation

The splitting of parallel corpora in terms of number of parallel sentences is shown in Table 4.1. Data is divided in training set, development set and test set. We use the training data to train the translation system and test set is used to confirm the results of the best method. Development set is used to optimize the model parameters for better translation quality. The parameters that are tuned using development set are weights for phrase translation table, language model, distortion model and weight for word penalty limit. Test set is left untouched during the training and development phase.

Corpus	Training Size	Development Size	Testing Size	Total Sentence Pairs
Emille	8,000	376	360	8,736
Penn Tree Bank	5,700	315	200	6,215
Quran	6,000	214	200	6,414
Bible	7,400	300	257	7,957

Table 4.1: Splitting of Parallel Corpora in terms of Sentence Pairs

[8] *Reordering probabilities will be learnt on phrases in both source and target directions.*

The data splitting follows the rule of taking the training sentences from the beginning of the corpora, followed by taking the sentences for the development set and the rest of the corpus is allocated to the test set.

Corpus	Training Size	Development Size	Testing Size	Total Sentence Pairs
Emille	141,136	6071	6,312	153,519
Penn Tree Bank	148,134	8,154	5,006	161,294
Quran	245,416	3,596	3,591	252,603
Bible	192,565	9,271	8,761	210,597

Table 4.2: Number of English tokens in our parallel corpora

Data splitting-summary in terms of number of tokens (words) for English chunk in parallel corpora is shown in Table 4.2 and for Urdu is shown in Table 4.3. Where, the numbers of words are based on full-form of the words including the punctuation marks.

Corpus	Training Size	Development Size	Testing Size	Total Sentence Pairs
Emille	183,016	8,322	8,841	200,179
Penn Tree Bank	169,539	9,934	6,216	185,689
Quran	262,124	3,805	4,061	269,990
Bible	186,175	9,349	8,403	203,927

Table 4.3: Number of Urdu tokens in our parallel corpora

4.2 Evaluation Measures

One of the most difficult tasks in Machine Translation is to evaluate the output of the system. For this study we have selected the BLEU (Bilingual Evaluation Understudy) (Papineni, et al., 2002) as an evaluation metric. The Bleu metric is an IBM-developed metric and very well known for the machine evaluation for the machine translation. It checks how closer the candidate translation is to the reference translation based on the n-gram comparison between both translations. The Bleu score is based on the number of correct n-gram matches between candidate and reference translation, and these matches are position-independent.

The Bleu metric ranges from 0 to 1. If the candidate translation is identical to the reference translation it will attain the score 1 and 0 in case of no similarities. Bleu metric is based on the modified n-gram precision measure for comparing the candidate translation against multiple reference translations.

$$Precision = \frac{Number\ of\ words\ from\ the\ candidate\ that\ are\ found\ in\ the\ reference}{Total\ number\ of\ words\ in\ the\ candidate} \quad 4.1$$

The metric modifies simple precision since MT system can over generate reasonable words, resulting in implausible, but high-precision, translations like Example 4.1 (Papineni, et al., 2002) below:

Example 4.1:

Candidate:	the the the the the the the.
Reference 1:	The cat is on the mat.
Reference 2:	There is a cat on the mat.

All of the seven words in the candidate translation appear in both reference translations, thus the candidate text is given the unigram accuracy that is shown in Equation 4.2.

$$Unigram\ Precision = \frac{7}{7} = 1 \quad 4.2$$

Now, for modified unigram precision calculation, for each word in the candidate translation, Bleu calculates its maximum total count in any of the reference translations. So in the Example 1 above, “the” appears twice in reference 1 and once in reference 2 so it’s MaxCount = 2. Now the total count of each word (Wc) in the candidate translation that is 7 for “the” in our example, is clipped to its MaxCount. Wc is then summed over all the words in the candidate translation.

$$Modified\ Unigram\ Precision = \frac{2}{7} = 0.28 \quad 4.3$$

Brevity penalty is introduced in the metric to penalize the shorter translations to receive too high score. Let, c be the length of the candidate translation and r be the effective reference corpus length. The brevity penalty (BP) is computed by,

$$BP = \begin{cases} 1 & if\ c > r \\ e^{(1-r/c)} & if\ c \leq r \end{cases} \quad 4.4$$

The final Bleu score is calculated by computing the geometric average of the modified n-gram precision, p_n using n-grams up to length N and positive weights w_n summing up to 1.

$$BLEU = BP\ .\exp(\sum_{n=1}^{N} w_n\ \ logp_n) \quad 4.5$$

While it is better to use several independent reference translations (usually 4 if available), our English-Urdu parallel data contain only 1 reference translation per sentence.

4.3 Types of Experiments

Various experiments are performed during this study for obtaining the Urdu translation from the given English sentence. We start with the baseline experiments followed by the experiments to observe the effect of a variety of improvement techniques that are applied to get the better translation quality. Experiments are performed on all four parallel corpora collected for this study. Parallel corpora domains and statistics are provided in detail in Section 2.1 and 2.3 respectively.

The main categories of experiments performed in this study are the following.

- Baseline Experiments
- Experiments with Distance-based reordering
- Experiments after applying word order transformation heuristic
- Experiments using factored-based translation
- Experiments with the combination of the techniques

Experiments are performed using the parallel data for training set, development data set and test set from same corpora, unless stated otherwise.

4.3.1 Baseline Experiments

Our baseline setup is a plain phrase-based translation model (i.e. single-factored) with only the bidirectional reordering model. In all experiments, language model consists of monolingual data and target Urdu corpora. To obtain the baseline results we perform different sets of experiments that are defined as follows.

i. Un-normalized target data with un-normalized language model.

ii. Normalized target data with normalized language model.

iii. Normalized target data with mix[9] language model.

[9] *Mix language model refers to the combination of un-normalized monolingual text and normalized target Urdu. Whereas un-normalized language model is combination of un-normalized monolingual data and un-normalized target corpora and normalized language model is combination of normalized monolingual data and normalized target corpora.*

iv. In all experiments where normalized target corpus is used, all Urdu data have been normalized, i.e. training data and reference translations of development and test data. Normalization steps are briefly discussed in Section 2.4.

Un-normalized Target Data with Un-normalized LM

First baseline experimental setup includes translation between source English data and un-normalized target Urdu data together with un-normalized language model. In Table 4.4 we present the results of the baseline experiment. The results are composed of BLUE score evaluated over the test corpora and the n-gram precisions of the trained system.

Parallel data	**BLEU-4**	***n*-gram precisions**							
		1	**2**	**3**	**4**	**5**	**6**	**7**	**8**
Emille	21.16	0.55	0.26	0.15	0.1	0.07	0.05	0.03	0.02
Penn Treebank	18.47	0.59	0.26	0.12	0.06	0.03	0.02	0.009	0.005
Quran	13.08	0.54	0.2	0.1	0.05	0.03	0.02	0.008	0.005
Bible	8.88	0.47	0.14	0.05	0.02	0.008	0.004	0.002	0.001

Table 4.4: Results of baseline system, with un-normalized target data and un-normalized language model

The table clearly demonstrates that it is more difficult to reproduce the reference translation of Bible than in the case of the other corpora. In Table 4.5 we show input sentence from the Bible corpus, its reference translation and its respective output translation obtained using the first baseline experimental settings.

Input	And death and hell were cast into the lake of fire. This is the second death.
Reference	پھِر مَوت اورعالمِ ارواح آگ کی جھِیل میں ڈالے گئے ۔ یہ آگ کی جھِیل دُوسری مَوت ہے ۔
Transliteration	phir maut aur ˀālmi arwāḩ āg ke jhīl meñ ḍāle gae. yah āg kī jhīl dūsrī maut he.
Output	اورمَوت اورجہنّم پھینک دِیا جائے جھِیل میں آگ کی ہے ۔ یہ ہے کہ دُوسری مَوت ہے ۔
Transliteration	aur maut aur jahaddam phīnk diyā jāe jhīl meñ āg kī he – yah he kah dūsrī maut he.

Table 4.5: Output translation of baseline system, with un-normalized target data and un-normalized language model

There are few issues associated with the translation generated by the baseline system. Firstly, although word جہنّم is also a correct translation of word "hell" besides عالمِ ارواح but it is wrong translation based on the context of the reference sentence. Secondly, the word "cast into" is wrongly translated into "پھینک دِیا جائے" which is actually the translation of "throwing something". The correct translation of "cast into" is "ڈالے گئے" that has meaning of "putting into".

Another major issue with baseline translation is the wrong syntactic ordering of phrases/words. We can see in Example 4.2 taken from Table 4.5 that the baseline system is unable to model the translation between language-pair that have different word order structures.

Example 4.2:

Input:	into the lake of fire
Reference:	آگ کی جھِیل میں
Transliteration:	āg kī jhīl meñ
Output:	جھِیل میں آگ کی
Transliteration:	jhīl meñ āg kī

Urdu uses reverse word order w.r.t. English in phrases of the type "X of Y". The reference word order in this example is "fire of lake into" whereas the translation by the baseline system has "lake into fire of". Although the system was able to flip "of fire" to "fire of", it failed to reorder the whole trigram correctly. That shows the necessity of including some notion of syntax in phrase-based MT systems. This problem can be further refine with the use of POS tagged language model that would ensure that the word order in reference phrase pair NN CM NN PR is mostly likely to occur compare to word order NN PR NN CM. where NN refers to “fire” and “lake”, CM refers to “of” and PR refers to “into”.

Normalized Target Data with Normalized LM

In the second baseline experimental setup, we perform translation between source English data and normalized target Urdu data together with normalized language model. In Table 4.6 we present the evaluation results of translations generated using current experimental setup.

Parallel data	BLEU-4	*n*-gram precisions							
		1	2	3	4	5	6	7	8
Emille	21.89	0.56	0.27	0.16	0.1	0.07	0.05	0.03	0.02
Penn Treebank	18.48	0.59	0.26	0.13	0.06	0.03	0.02	0.008	0.005
Quran	14.02	0.54	0.21	0.01	0.05	0.03	0.01	0.007	0.003
Bible	9.10	0.5	0.15	0.05	0.02	0.007	0.002	0.001	0.00

Table 4.6: Results on baseline system, with normalized target data and normalized language model

In Table 4.7, the input sentence from Table 4.5 is presented to the baseline system trained on normalized data. The target reference sentence has been normalized, too.

Input	And death and hell were cast into the lake of fire. This is the second death.
Reference	پھِر مَوت اورعالِم ارواح آگ کی جھِیل میں ڈالے گئے ۔ یہ آگ کی جھِیل دُوسری مَوت ہے ۔
Transliteration	phir maut aur ˀālmi arwāḩ āg ke jhīl meñ ḍāle gae. yah āg kī jhīl dūsrī maut he.
Output	اورموت اورعالم ارواح تھے میں ڈال دیا ۔ کی جھیل کے آگ ہے ۔ یہ ہے کہ دوسری موت ہے ۔
Transliteration	aur maut aur ˀālm arwāḩ the meñ ḍāl diyā. kī jhīl ke āg he. yah he kah dūsrī maut he.

Table 4.7: Output translation of baseline system, with normalized target data and normalized language model

We can see few improvements in the obtained translation in Table 4.7 as compare to the output translation in Table 4.5. The improvements are correct translation of word “hell” and also “cast”. But conversely the output translation using current experimental settings has more un-necessary words compare to the un-normalized translation scheme presented in Table 4.5. If we remove additional words “تھے”, “کے” and “ہے” from the output translation and reorder the verb phrase and noun phrase then the output translation can be understandable.

Normalized Target Data with Mixed LM

In Table 4.8 we present the evaluation results of baseline experiments performed on source English data and target normalized Urdu data together with mixed LM.

Parallel data	**BLEU-4**	***n*-gram precisions**							
		1	**2**	**3**	**4**	**5**	**6**	**7**	**8**
Emille	21.61	0.54	0.26	0.15	0.1	0.07	0.05	0.03	0.02
Penn Treebank	18.54	0.6	0.27	0.13	0.06	0.03	0.02	0.009	0.006
Quran	13.14	0.56	0.21	0.1	0.05	0.02	0.01	0.007	0.004
Bible	9.39	0.5	0.15	0.05	0.02	0.008	0.004	0.002	0.00

Table 4.8: Results on baseline system, with normalized target data and mixed language model

In Table 4.9, the input sentence from Table 4.5 and Table 4.7 is presented to the baseline system trained on normalized target corpus and mixed LM. The target reference sentence has been normalized, too. The output translation in Table 4.9 is roughly similar to the translation in Table 4.7 with other additional words.

Input	And death and hell were cast into the lake of fire. This is the second death.
Reference	پھِر مَوت اورعالمِ ارواح آگ کی جھِیل میں ڈالے گئے ۔ یہ آگ کی جھِیل دُوسری مَوت ہے ۔
Transliteration	phir maut aur ˀālmi arwāḩ āg ke jhīl meñ ḍāle gae. yah āg kī jhīl dūsrī maut he.
Output	اورموت اورعالم ارواح تھے اس میں ڈالا کی جھیل کے آگ ہے ۔ یہ ہے کہ دوسری موت ہے ۔
Transliteration	aur mwt awr ˀālm arwāḩ the as mīñ ḍālā kī jhīl ke āg he. yah he kah dūsrī mwt he.

Table 4.9: Output translation of baseline system, with normalized target data and mixed language model

To summarize all baseline experiment results, in Table 4.10 we compare BLEU scores before and after applying normalization on target corpora and language model.

Parallel data	**BLEU Score**		
	Un-normalized Urdu Corpus / Un-Normalized LM	**Normalized Urdu Corpus / Normalized LM**	**Normalized Urdu Corpus / Mixed LM**
Emille	21.16	21.89	21.61
Penn Treebank	18.47	18.48	18.54
Quran	13.08	14.02	13.14
Bible	8.88	9.10	9.39

Table 4.10: comparison of baseline experiment results

From Table 4.10 we can see that the BLUE score using un-normalized settings is always less than the other two normalized experimental settings in comparison to the results of all the corpora. Although the difference in BLEU score is not very significant in both un-normalized and normalized settings but our assumption for the gain in BLUE score for normalized target corpora is that the normalization helps in improving the translation model. Words that can be written in multiple ways are now written the same way in both training and test data, which makes it easier to learn the translation. This reason was also the motivation behind normalizing the Urdu data. However, we don't claim that normalized settings always work better than the un-normalized settings and hence this observation is further required to be explored.

The evaluation results of normalized LM and mixed LM experimental settings have some random behavior. The Quran corpus has significant rise in BLEU score using normalized LM settings as compare to other two settings. The most apparent reason of this improvement is the large amount of Islamic monolingual data used in building LM that helps in improving the translation of Quran data. Penn and Bible data has small improvement over mixed LM settings compare to normalized LM.

Mixed LM brings (mostly) better results than the other configurations. The reason could be that phrases that occur in the phrase table are covered by the LM (in the same form, i.e. if the parallel corpus is normalized, its Urdu part is included in LM also normalized). However, normalizing the rest of the monolingual data (which is much larger) probably just removes the information, while it has less direct impact on phrases from the parallel corpus. So far it's just a hypothesis, because we did not have time to collect supporting evidence and the chances could be that the deviation in BLEU score is merely random because the BLEU score drop does not seem to be statistically significant.

The Mixed language model setting was initially created unintentionally but after seeing the results we decided to use its experimental settings with the all of the remaining experiments. Also, for comparisons among results using different experimental setup, we use the baseline results of normalized target data and mixed language model.

4.3.2 Experiments with Distance-Based Reordering

In this section we perform the experiments using the distance-based reordering model together with the bidirectional orientation model. The experiments are performed using the default distortion-limit defined in Moses. In Table 4.11, we show the results after using the distance-based reordering model on source and normalized target data.

Parallel data	**BLEU-4**	***n*-gram precisions**							
		1	**2**	**3**	**4**	**5**	**6**	**7**	**8**
Emille	23.59	0.57	0.29	0.17	0.11	0.08	0.05	0.04	0.03
Penn Treebank	22.74	0.6	0.3	0.17	0.09	0.05	0.03	0.02	0.01
Quran	13.99	0.55	0.2	0.1	0.05	0.02	0.01	0.009	0.005
Bible	13.16	0.5	0.19	0.08	0.04	0.02	0.01	0.004	0.001

Table 4.11: Results of Distance-based reordering on source and normalized target data

There is a significant rise in BLEU score of experiments with distance-based reordering as compared to the baseline experiments results. That doesn't necessarily indicate improvement in translation quality, as correlation between BLEU and human judgments is known to be lower for inflectional free-word-order languages. Thus to verify also the improvement in translation quality in Table 4.12 we manually analyze the output of the distance-based system on the previously discussed input sentence from Bible data.

Input	And death and hell were cast into the lake of fire. This is the second death.
Reference	پھِر مَوت اورعالمِ ارواح آگ کی جھِیل میں ڈالے گئے ۔ یہ آگ کی جھِیل دُوسری مَوت ہے ۔
Transliteration	phir maut aur ˀālmi arwāḥ āg ke jhīl meñ ḍāle gae. yah āg kī jhīl dūsrī maut he.
Output	تھے اوران کی موت اورعالم ارواح آگ کی جھیل میں ڈالا جاتا ہے ۔ یہ دوسری موت ہے ۔
Transliteration	the aur un kī maut aur ˀālm arwāḥ āg kī jhīl meñ ḍālā jātā he. yah dūsrī maut he.

Table 4.12: Output translation after adding Reordering Model

Hence, after adding the reordering model we can see in output translation the correct ordering of phrase pair "into the lake of" which is previously discussed in Example 4.2. Also the verb phrase is precisely preceded by the objectival phrase " آگ کی جھیل میں". There are still two major problems left with the obtained translations. Firstly, the un-necessary phrase "تھے اور ان کی" at the beginning of the sentence that makes the translation difficult to understand and secondly the wrong case ending of the verb phrase.

Although output translation from the system with reordering model is not very good but, the reordering of the words at least makes quite rational word ordering in the output translation compared to the translation produced by the baseline system. Also, the translation of distance-based system is roughly understandable but output translation of baseline system is not understandable at all.

4.3.3 Experiments after Applying Word Order Transformation

We further performed experiments with preprocessed source corpora i.e. reordered English data using word order transformation scheme. In this experiment we only use the bidirectional orientation model of Moses. The results of experiments are presented in Table 4.13.

Parallel data	BLEU-4	*n*-gram precisions							
		1	2	3	4	5	6	7	8
Emille	25.15	0.56	0.3	0.19	0.13	0.1	0.07	0.05	0.04
Penn Treebank	24.07	0.6	0.3	0.18	0.1	0.06	0.03	0.02	0.01
Quran	13.37	0.5	0.2	0.09	0.05	0.03	0.01	0.007	0.002
Bible	13.24	0.5	0.18	0.08	0.04	0.02	0.008	0.003	0.001

Table 4.13: Translation Results after applying word order transformation scheme

In Table 4.14 we compare the BLEU scores of baseline, distance-based model and word order transformation scheme. The results show the significant improvement in BLEU score of transformation-based model over the baseline and distance-based reordering model. Except in the Quran data where translation accuracy has decreased from 13.99 to 13.37 compare to the distance-based model. One potential reason of drop in BLEU score could be atypical long sentences in Quran data while our transformation system contains limited number of transformation rules for reordering the long sentences.

Parallel data	BLEU Score		
	Baseline	Distance-based Model	Word order Transformation Model
Emille	21.61	23.59	25.15
Penn Treebank	18.54	22.74	24.07
Quran	13.14	13.99	13.37
Bible	9.39	13.16	13.24

Table 4.14: Comparison of baseline, distance-based model and transformation-based model Results

In Table 4.15 we present the previously discussed input sentence from Bible data and its output translation generated by our reordering system.

Input	And death and hell were cast into the lake of fire. This is the second death.
Reference	پھِر مَوت اورعالمِ ارواح آگ کی جھِیل میں ڈالے گئے ۔ یہ آگ کی جھِیل دُوسری مَوت ہے ۔
Transliteration	phir maut aur ˀālmi arwāḩ āg ke jhīl meñ ḍāle gae. yah āg kī jhīl dūsrī maut he.
Output	اورموت اورعالم ارواح آگ کی جھیل میں ڈالا تھا ۔ یہ دوسری موت ہے ۔
Transliteration	phr’ maut aur ˀālm arwāḩ āg kī jhīl meñ ḍāle gae. yah āg kī jhīl dūsrī maut he.

Table 4.15: Output translation after preprocessing English data

The reordering problem in phrase pair that is previously discussed in Example 4.2 is correctly translated into the output translation of transformation model. The ordering of subject, object and verb phrase is also correctly transformed into the default Urdu sentence structure. The interesting observation in this translation is the correct translation of word “and” at the beginning of the sentence onto the word “اور” unlike the translation generated by distance-based models and baseline systems. Although “and” has the translation “پھر” in reference translation but that is actually translation of “then” in Urdu not “and”. By keeping this fact in mind, our system has generated the translation that makes the sentence relatively understandable. The problem with the verb case ending also exists in our system.

In Table 4.16 we present an input sentence from the Penn-Treebank corpus together with the translation output from different systems. Here we would like to mention that the reference translation of given sentence is not a well structured sentence. Reference is splitted into two sub clauses separated with the comma where the better translation could be written using a single clause like input sentence. Distance-based system tries to perform the reordering within certain word limits (e.g. the default of 6 words) whereas our transformation output has applied the long distance word ordering by correctly taking the main verb phrase at the end of the sentence.

Original sentence	The Senate Banking Committee will begin hearings next week on their proposal to expand existing federal housing programs.
Transformed input	The Senate Banking Committee hearings next week their proposal existing federal housing programs expand to on begin will.
Reference	سینیٹ بنکنگ کمیٹی سماعتیں اگلے ہفتے شروع کرے گی، موجودہ وفاقی ہاؤسنگ پروگراموں کو وسیع کرنے کی ان کی تجویز پر۔
Transliteration	sīneṭ banking kamīṭī smāʾateñ agale hafte šurūʾ kare gī, mūjūdah waqāfī hāusing progrāmoñ ko wasīʾa karne kī un kī tajwīz par.
Baseline	سینیٹ بنکنگ کمیٹی شروع کرے گی hearings اگلے ہفتے کے طورپر ان کی تجویز کو وسیع کرنے کے لیے موجودہ وفاقی ہاؤسنگ پروگراموں کے ۔
Transliteration	sīneṭ banking kamīṭī šurūʾ kare gī hearings agale hafte ke ṭūr par un kī tajwīz ko wasīʾa karne ke līe mūjūdah waqāfī hāusing progrāmoñ ke.
Distance model	سینیٹ بنکنگ کمیٹی اگلے ہفتے شروع کرے گی ان کی تجویزپر hearings موجودہ وفاقی ہاؤسنگ پروگراموں کے وسیع کرنے کے لیے ہے ۔
Transliteration	sīneṭ banking kamīṭī agale hafte šurūʾ kare gī un kī tajwīz par hearings mūjūdah waqāfī hāusing progrāmoñ ke wasīʾa karne ke līe he.
Transformation scheme	سینیٹ کی بنکاری کمیٹی سماعتیں اگلے ہفتے ان کی تجویزپر موجودہ وفاقی ہاؤسنگ پروگراموں کے وسیع کرنے کے لیے پر شروع کرے گی ۔
Transliteration	sīneṭ kī bankārī kamīṭī smāʾateñ agale hafte un kī tajwīz par mūjūdah waqāfī hāusing progrāmoñ ke wasīʾa karne ke līe par šurūʾ kare gī.

Table 4.16: Output translations after applying word order transformation

The other noticeable fact is the correct translation of object phrase “hearings” by our system whereas the less sophisticated systems were unable to translate the object noun phrase. The plausible reason of translation of “hearings” is the formation of phrase pair “The Senate Banking Committee hearings” by our system which also exists in the training corpus. Thus this phrase construction helped in retrieving correct translation of “hearings” from phrase table.

In Urdu, constituents of compound noun phrases in the form “NNP_1 NNP_2” are separated using postpositions “NNP_1 IN NNP_2”. Due to brining subject and object phrase closer, much better translation of subject phrase (consists of compound noun) as shown in Example 4.3 is retrieved as appose to using its transliterated form as used in reference sentence.

Example 4.3:

Input:	Senate NNP_1	Banking NNP_2	Committee NNP_3	
Reference:	کمیٹی NNP_3	بنکنگ NNP_2	سینیٹ NNP_1	
Output:	کمیٹی NNP_3	بنکاری NNP_2	کی IN	سینیٹ NNP_1

According to our analysis the output translation produced by transformation system is much accurate then the output produced by baseline and distance-based models except the additional postposition “پر” before the verb phrase “ شروع کرے گی” at the end of the sentence. The reason of placing this postposition before verb phrase is quite obvious because of the incorrect occurrence of preposition “on” before verb phrase “begin will” in transformed input sentence.

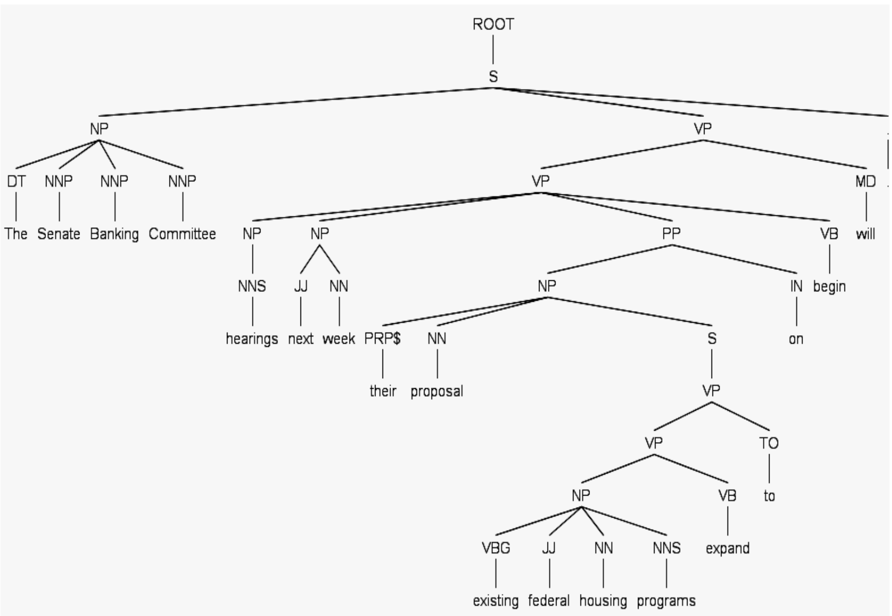


Figure 4.1: Transformed English tree of input sentence presented in Table 4.16.

In Figure 4.1 we show the reason of incorrect placing of preposition “on” before verb phrase. In our transformed tree the transformation rule PP -> IN NP correctly transformed into PP -> NP IN but this transformation actually generated error in the output translation because of the existence of sub-phrase “S” inside the noun phrase (NP). After deep analysis of sub-phrase existence, we found out that in all those sentences where sub-phrases exist in the form of “S” or “SBAR” (notions of Stanford Parser) we could programmatically remove the sub-phrase node and place it at the end of current transformation rule. For instance in our case the rule PP ->NP IN will become PP -> NP IN S in transformed tree. The same scheme is also applicable for several other cases where sub-phrases split the constituents of phrase pair and cause error in translation. The current transformation system doesn’t include the proposed sub-phrasal techniques and we can produce more sophisticated translation output by our system after applying the sub-phrasal translation scheme.

Due to syntactic reordering, the system has resulted into producing better translation output not only compare to a baseline systems but also distance-based models and can be improved further by applying the proposed changes.

4.3.4 Experiments with Factored-Based Model

In this section we perform the experiments using advance translation system of (plain) phrase-based MT i.e. factor-based model of Moses. The major reason of using factor-based model is to overcome the data sparseness issue that occurs due to translating the highly inflectional languages. In the following experiments we only use the bidirectional orientation model for reordering the phrases.

We tried three different experimental settings in factor-based models as defined below.

i. Array of factors compose of word, lemma and part-of-speech tag on both source and target side.

ii. Only word and lemma on source side and word, lemma and part-of-speech on target side.

iii. Only single factor (word) on source side and two factors i.e. word and part-of-speech tag on target side.

Due to extensive memory requirements by the complex factor models in experiment (i) and (ii), we were unable to build the translation model. Whereas we only succeeded in building translation model using experimental setting (iii) which is less complicated than other two models. In all experiments based on factor-based model we use simple factor model i.e. experimental setting (iii) together with part-of-speech tagged language model. In Table 4.17 we present the baseline results and the factorization results that are achieved after performing the experiments by using factor-based model.

Parallel data	BLEU-4	
	Factorization	Baseline
Emille	17.48	21.61
Penn Treebank	16.92	18.54
Quran	12.92	13.14
Bible	8.55	9.39

Table 4.17: Translation Results of using only Factor-based model

As we can see from Table 4.17 that with the use of only factorization, BLEU score has decreased significantly compare to the baseline results. We further try experiments using factorization together with the distance-based reordering model and also transformation-based reordering model. In Table 4.18 and Table 4.19 we further provide the evaluation results of experiments performed simultaneously with distance-based reordering model and transformation-based reordering model together with the factorization model.

Parallel data	BLEU-4	
	Factorization + Distance Based Model	Only Distance-Based Model
Emille	23.35	23.59
Penn Treebank	19.82	22.74
Quran	12.62	13.99
Bible	12.25	13.16

Table 4.18: Translation Results of using factorization with Distance-based reordering model

As we can see from Table 4.18 and Table 4.19 that adding factorization doesn't bring the significant improvement over already achieved results by using only reordering models but it indeed improved the results obtained in Table 4.17 by using only factorization without distance-based and transformation-based rendering schemes. In Table 4.20 we further investigate the output translation using factored-based model and the possible reason of the decrease in BLEU score.

Parallel data	BLEU-*n* 4	
	Factorization + Transformation Based Model	Only Transformation-Based Model
Emille	25.18	25.15
Penn Treebank	22.64	24.07
Quran	13.59	13.37
Bible	11.86	13.24

Table 4.19: Translation Results of using factorization with Transformation-based reordering model

In the example shown in Table 4.20 we can see that the verb “use” is translated into pos tag sequence “NN VBL AUXT SM” and positioned at the start of the sentence instead of putting it at the end of the sentence. The pos tag sequence “PR KER NNCM QW AUXA AUXT SM” at the end of the sentence don’t correspond to any word pair in the input sentence and also don’t match with the reference sentence. The noticeable point here is that in reference translation verb phrase “استعمال” is wrongly tagged as noun instead of verb and verb “کیجئے” is also tagged as noun instead of light verb.

Input	Use this leaflet as a guide.
Reference	اس|PR لیف|NN لٹ|NN کو|CM بطور|RB گائیڈ|NNP استعمال|NN کیجئے|NN ۔|SM
Transliteration	us|PR līf|NN laṭ|NN ko|CM baṭor|RB gāīḍ|NNP astaˀamāl|NN kījie|NN .|SM
Output	استعمال|NN کرتے|VBL ہیں|AUXT ۔|SM اس|PR لیفلٹ|NN کو|CM گائیڈ|NNP کے|KER طور|NN پر|CM اس|PR کے|KER بعد|NNCM کیا|QW ہوتا|AUXA ہے|AUXT ۔|SM
Transliteration	astaˀamāl|NN karte|VBL heñ|AUXT .|SM us|PR līflaṭ|NN ko|CM gāīD|NNP ke|KER ṭor|NN par|CM us|PR ke|KER bˀd|NNCM kyā|QW hotā|AUXA he|AUXT .|SM

Table 4.20: Output translation using Factored-based model

After analyzing the multiple output translations of factorization model only we found out that due to lack of reordering, factorization model produce the translation of verbal phrase in the beginning or middle of the sentence by following the word order of input sentence and also try to embed extra auxiliaries or verbal phrase at the end of sentence by following the probable POS tag sequence of Urdu sentence i.e. SOV using the POS tagged LM.

Input	Use this leaflet as a guide.
Transformed input	A guide as this leaflet use.
Reference	اس|PR لیف|NN لٹ|NN کو|CM بطور|RB گائیڈ|NNP استعمال|NN کیجئے|NN ۔|SM
Transliteration	us|PR līf|NN laṭ|NN ko|CM baṭor|RB gāīḍ|NNP astaˀamāl|NN kījie|NN .|SM
Output	اس|PR لیف|NN لٹ|NN کو|CM بطور|RB گائیڈ|NNP استعمال|NN کریں|VB ۔|SM
Transliteration	us|PR līf|NN laṭ|NN ko|CM baṭor|RB gāīḍ|NNP astaˀamāl|NN kareñ|VB .|SM

Table 4.21: Output translation using Factored-based model and Transformation-based reordering

We can see the improved translation (after using the transformation model together with the factorized system) of the same sentence presented previously in Table 4.20. The only difference in reference and output translation is the different form of the verb phrase i.e. translation of "use" into "کریں" instead of "کیجئے". Due to marking of correct verb form as noun in reference translation, POS tagged LM give more significance to the phrasal pair tagged as verb to put at the end of the sentence. From this example we can see that the wrong POS tagging is also the cause in decrease of evaluation score even then the translation is quite understandable. We couldn't gather further evidence on decrease in BLEU score because of using factorization model and hence this dilemma is still need to be resolved.

Summary

In this chapter we performed different set of experiments to produce the output Urdu translation given the source English sentence. To refine the output generated by baseline systems we further carried out experiments using different reordering models and Moses factorized phrase-based MT. The output sentences generated using transformation-based reordering generally performed well over the output generated by distance-based reordering models. Moreover, factorized phrase-based MT didn't' bring improvement in evaluation results and same could be assumed for the translation quality as well. In this work we provided the initial hypothesis on the failure of factorization but this is only our assumption and it is further required to be explored.

Chapter

5

Discussion and Conclusion

In the preceding chapters we have seen the specific improvement techniques in the domain of statistical machine translation for English-Urdu language pair. The general idea was to produce the grammatically coherent and human understandable translation given the input English sentence. In this final chapter, we summarize our approach and substantiate key results. We further provide the comparison to related work, and we close this study work by drawing conclusions and giving directions of future research.

5.1 Summary

In this study, we address the translation issues between languages with significant word order differences modeled using the phrase-based machine translation systems. In order to approach the translation issues due to word order difference, we captured the syntactic structure of natural language by parsing the source English corpora.

We initiated this research work with the collection of English-to-Urdu parallel corpora and huge target side monolingual corpora. Then we further proceeded with the description of the translation issues inherent to the (simple) phrase-based machine translation systems and devised different techniques to improve the quality of the translation produced by PBT systems. Thus, the introduced techniques are based on modeling translation issues from two perspectives (i) dealing with the issues caused by difference in syntactic structure of distant word order languages, and (ii) introducing morphology into the system to overcome the data sparseness issue extremely probable for translating highly inflected languages.

The improvement techniques itself give rise to the further two questions that how syntax would be possibly integrated into the PBT systems and how we can formulate the model that deal with the data sparseness problem. We tackled with the first problem by reducing word order difference between both languages i.e. made both languages syntactically similar to each other. We parse a source English corpus and apply the word order transformation over a corpus. This results in transformed English corpus having the syntactic structure similar to the structure of target side of parallel corpora. The transformation is applied after the deep analysis of parallel corpora and by extracting the transformation rules that represent the most common word order mapping of syntactic structures. This technique indeed showed its viability, leading to a potential improvement in translation accuracy in terms of BLUE score for instance for Emille Corpus from 21.61% to 25.15% compared to our baseline system from 23.6% to 25.15% compare to the distance-based reordering model of PBT systems.

The second problem is dealt by using the factored-based translation systems which is an extended framework of PBT systems. The factored-based models overcome the data sparseness issue by using the additional part-of-speech tagged language model that generalizes well over the n-grams that are not seen before by using the correct tag sequences but possibly with the different words. We first tried to build the complex factorization model but couldn't succeed due to extensive memory requirement. Further we continued using the simple factorized model which didn't provide the satisfactory results on baseline experiments but equally performed well together with the use of transformation-based reordering model. However, there is a potential space for improvement by using the more accurate POS tagger for Urdu. Our translation improvement approaches can show more capabilities if we can add further training data into the system (current systems are trained on only few thousands of parallel sentences).

In the following section we compare our approach to the related work to our study i.e. with Google's English-to-Urdu Statistical Machine Translation System.

5.2 Comparison to Related work

In this section we compare our evaluation scores with the Google's English to Urdu translation system with our four translation systems trained on Emille, Penn Treebank, the Bible and the Quran data. In Table 5.1 we present the evaluation scores on translation produced by Google's translation system on the same normalized test data used for the evaluation of our trained systems and its comparison with the results of output produced by our baseline system and word order transformation system.

Parallel data	BLEU-4		
	Google's Translation System	Our System Baseline	Transformation System
Emille	19.31	21.61	25.15
Penn Treebank	9.30	18.54	24.07
Quran	21.44	13.14	13.37
Bible	5.72	9.39	13.24

Table 5.1: comparison of the results produced by Google's translation system and our baseline and word order transformation system

Our transformation systems clearly outperformed Google's evaluation scores on three of the parallel data results except Quran data. Although we cannot directly compare both systems BLEU scores as both systems are trained on different parallel data. Nevertheless, Google systems use parallel data that consist of millions of tokens perhaps collected from various domains and our systems are trained on few thousand of parallel sentences extracted from limited domains. This fact can indirectly lead us to the comparison of both systems output.

After having compared our work to related research work, we move to the last part of this chapter, concluding and pointing out directions of future work.

5.3 Conclusion and future work

In the presented text, we have described improvements in English-Urdu translation produced using phrase-based machine translation system, Moses. We applied two techniques where we achieved significantly improved results after applying preprocessing technique on source data. The results obtained after applying preprocessing on data can be improved further by applying the proposed modifications in word order transformation system. We are looking forward to further improve this work by possibly introducing the bilingual dictionary to minimize the percentage of out of vocabulary words that remain un-translated in the system output. In future we would like to further improve the reordering model of transformation system by accumulating more transformation rules that covers relatively long sentences as well.

English Grammar Rule	Transformation Order
S -> NP VP	nochange
S -> ADVP VP	reverse
S -> ADVP VP NP	2 0 1
SBAR -> WHNP S	nochange
SINV -> ADVP VP NP	2 0 1
SINV -> MD NP VP	1 2 0
SQ -> MD NP VP	1 2 0
SQ -> VB* RB NP VP	nochange
NP -> NP PP	reverse
NP -> NP PP PP	0 2 1
NP -> NP PP .	1 0 2
NP -> DT NN RB	2 0 1
NP -> DT NN S	2 0 1
NP -> NP NN NNS	2 1 0
NP -> NP PRN PP	2 0 1
NP -> NP LRB PP RRB	0 3 2 1
NP -> RB JJ PRN	2 0 1
NP -> default	nochange
VP -> TO VP	reverse
VP -> VB* NP	reverse
VP -> VB* PP	reverse
VP -> VB* NP UCP	1 0 2
VP -> VB* ADJP	reverse
VP -> VB* ADVP	reverse
VP -> VB* ADVP ADVP	1 2 0
VP -> VB* S	nochange
VP -> VB* : S	nochange
VP -> VB* S : S	nochange
VP -> VB* : SQ	nochange
VP -> VB* PP : SQ	1 0 2 3
VP -> VB* ADJP	nochange
VP -> VB* ADJP , ADJP	1 2 3 0
VP -> VB* ADVP VP	1 2 0
VP -> ADVP VB* NP	0 2 1
VP -> ADVP VB* PP	2 0 1
VP -> ADVP VB* PP SBAR	2 0 1 3
VP -> VB* RB VP	2 0 1
VP -> MD RB VP	2 0 1
VP -> MD ADVP VP	1 2 0
VP -> MD , ADVP , VP	nochange
VP -> MD RB ADVP VP	2 3 0 1
VP -> MD RB PP VP	2 3 0 1

VP -> VB* NP PP	reverse
VP -> VB* PP NP	1 2 0
VP -> VB* PP NP S	1 2 0 3
VP -> VB* NP PP PP	1 2 3 0
VP -> VB* PP PP	1 2 0
VP -> VB* PP PP , SBAR	1 2 0 3 4
VP -> VB* NP NP	1 2 0
VP -> VP CC VP	nochange
VP -> ADVP VP CC VP	nochange
VP -> VP , CC VP	nochange
VP -> VP , VP CC VP	nochange
VP -> VP CC VP CC VP	nochange
VP -> VP , CC VP PP	nochange
VP -> , CC VP :	nochange
VP -> VB* CC VB* NP	0 1 3 2
VP -> VP , NP	2 1 0
VP -> VB* , PP , S	1 2 3 0 4
VP -> VB* ADJP S	nochange
VP -> VB* ADJP S SBAR PP	1 0 2 3 4
VP -> VB* ADVP PP	1 2 0
VP -> VB* SBAR	nochange
VP -> VB* PRN	nochange
VP -> VB* PRN SBAR	nochange
VP -> VB* PP SBAR	1 0 2
VP -> VB* RB ADJP SBAR	2 1 0 3
VP -> VB* NP ADVP	1 2 0
VP -> VB* NP ADVP SBAR	1 2 0 3
VP -> VB* NP ADVP PP	1 3 2 0
VP -> VB* NP ADVP PP SBAR	1 3 2 0 4
VP -> ADVP VP NP ADVP	3 2 0 1
VP -> VB* PRT NP SBAR	nochange
VP -> VB* NP PRT PP PP	2 3 4 1 0
VP -> VB* PRT NP ADVP , SBAR	3 2 1 0 4 5
VP -> VB* ADVP ADJP S	1 2 0 3
VP -> RB VP CC ADVP VP	nochange
VP -> VB* NP PP , CC VB* NP	2 1 0 3 4 6 5
VP -> default	reverse
PP -> IN NP	reverse
PP -> TO NP	reverse
PP -> IN S	reverse
ADJP -> default	nochange
ADVP -> ADVP PP	reverse
ADVP -> RBR IN RB	reverse
WHPP -> IN WHNP	reverse

B.1. Source Sentences

1) you can get these from your social security office .
2) poor transport contributes to social exclusion in two ways .
3) first , it restricts access to activities that enhance people ' s life chances , such as work , learning , health care , food shopping , and other key activities .
4) second , deprived communities suffer disproportionately from pedestrian deaths , pollution and the isolation which can result from living near busy roads .
5) there are number of contributors to social exclusion .
6) poor transport is just one of them .
7) many people experiencing social exclusion will not suffer from poor transport .
8) however , poor transport can be an important factor in restricting access to opportunity .
9) it can therefore undermine key government objectives on welfare to work , raising educational achievement and narrowing health inequalities , and has costs for individuals , businesses , communities and the state .
10) transport can be a significant barrier to accessing work :
11) two out of five jobseekers say lack of transport is a barrier to getting a job .
12) one in four jobseekers say that the cost of transport is a problem getting to interviews .
13) one in four young people have not applied for a particular job in the last 12 months because of transport problems .
14) one in 10 people in low - income areas have turned down a job in the last twelve months because of transport .
15) young people with driving licences are twice as likely to get jobs than those without .
16) poor transport is linked to young people dropping out of college :
17) sixteen - to 18 - year - olds spend on average £ 370 a year on transport .
18) forty - seven per cent of 16 - to 18 - year - olds experience difficulty with this cost .
19) six per cent of 16 - to 24 - year - olds turn down training or further education opportunities because of problems with transport .
20) for those who rely on public transport , getting to hospitals is particularly difficult , and can lead to missed health appointments :
21) thirty - one per cent of people without a car have difficulties travelling to their local hospital , compared to 17 per cent of people with a car .
22) seven per cent of people without cars say they have missed , turned down , or chosen not to seek medical help over the last 12 months because of transport problems .
23) this is double the rate in the general population .
24) children from the lowest social class are five times more likely to die in road accidents than those from the highest social class .
25) sixteen per cent of people without cars find access to supermarkets hard , compared with six per cent of people with cars .
26) poor transport can also affect people ' s participation in a range of other activities .
27) including seeing friends and family , volunteering and caring , religious activities , exercise and cultural activities .
28) eighteen per cent of people without a car find seeing friends and family difficult because of transport , compared with eight per cent for car owners .

29) people without cars are also twice as likely to find it difficult getting to leisure centres (nine per cent) and libraries (seven per cent) .
30) nearly one in three households does not have access to a car .
31) they depend primarily on walking to get around , but also on buses , lifts from family and friends and taxis .
32) cycling and rail make up a tiny fraction of their journeys .
33) 9 . people can face three types of barriers to accessing work , learning , health care and other key activities :
34) access and availability : people cannot get to key places in a reasonable time , reliably and safely .
35) this may be due to poor network coverage , frequency , and reliability of public transport or a lack of accessible facilities .
36) only 20 per cent of buses and 10 per cent of trains meet new accessibility regulations under the disability discrimination act .
37) in addition people living in rural areas without a car face particularly acute problems due to longer walking distances to bus stops , and low service frequency .
38) cost : people cannot afford personal or public transport .
39) bus fares have risen by nearly a third in the last fifteen years .
40) low - income households that do have a car spend nearly a quarter of their weekly household expenditure on motoring .
41) travel horizons : people are unwilling to travel long journey times or distances , or may lack trust in , or familiarity with , transport services .

B.2. Reference Translation

1) یہ آپ کو اپنے سوشل سکیورٹی آفس سے مل سکتے ہیں ۔
2) ناقص ٹرانسپورٹ سے سماجی اخراج) سماج سے اخراج ۔ سوشل ایکسکلوژن (میں دو طریقے سے اضافہ ہوتا ہے ۔
3) پہلے تو یہ کہ ، اس سے ان سرگرمیوں تک رسائی محدود ہو جاتی ہے جو لوگوں کی زندگی کے مواقع کی کیفیت میں اضافہ کرتی ہیں ، جیسے کام ، تعلیم حاصل کرنا ، صحت کی دیکھ بھال ، خوراک کی شاپنگ ، اوردوسری کلیدی سرگرمیاں ۔
4) دوسرے یہ کہ ، محروم کمیونٹیوں کو پیدل چلنے والوں کی اموات ، ماحول کی آلودگی اورمصروف سڑکوں کے قریب رہنے کی وجہ سے امکانی لگ تھلگ پن کی تکالیف کا غیرمتناسب حد تک سامنا کرنا پڑتا ہے ۔
5) سماجی اخراج میں اضافہ کرنے والے متعدد عوامل ہیں ۔
6) ناقص ٹرانسپورٹ ان میں سے صرف ایک ہے ۔
7) سماجی اخراج برداشت کرنے والے بہت سے لوگوں کو ناقص ٹرانسپورٹ کا سامنا نہیں کرنا پڑے گا ۔
8) تاہم ، ناقص ٹرانسپورٹ موقعہ تک رسائی کو محدود کرنے کا ایک اہم ذریعہ ہو سکتا ہے ۔
9) اس لئے ، یہ کام ، تعلیمی کامیابی کے درجہ کو بلند کرنے ، اورصحت کے اندرمختلف قسم کی نابرابری کوکم کرنے کے سرکاری مقاصد جن کا تعلق فلاح سے ہے پر منفی اثر ڈال سکتی ہے ، اوراس کی افراد ، کاروباری اداروں ، کمیونٹیوں اور ریاست کو بھاری قیمت ادا کرنی پڑتی ہے ۔
10) ٹرانسپورٹ کام تک رسائی کے راستے میں ایک قابل ذکر رکاوٹ ہو سکتی ہے:
11) پانچ روزگار تلاش کرنے والوں میں سے دو کا کہنا ہے کہ ناقص ٹرانسپورٹ کام تک پہنچنے کی راہ میں رکاوٹ ہے ۔

12) روزگار تلاش کرنے والے چار افراد میں سے ایک کا کہنا ہے کہ ٹرانسپورٹ کی لاگت انٹرویو میں پہنچنے کے سلسلے میں ایک مسئلہ ہے ۔

13) چار نوجوانوں میں سے ایک نے ٹرانسپورٹ کے مسائل کی وجہ سے پچھلے 12 مہینے میں کسی خاص جاب کے لئے درخواست نہیں دی ۔

14) کم آمدنی والے علاقوں میں 10 افراد میں سے ایک ایسا ہے جس نے پچھلے بارہ مہینے میں جاب لینے سے ٹرانسپورٹ کی وجہ سے انکار کیا ہے ۔

15) ڈرائیونگ لائسنس والے نوجوانوں کا لائسنس کے بغیر لوگوں کے مقابلے میں جاب حاصل کرنے کا امکان دوگنا ہے ۔

16) ناقص ٹرانسپورٹ کالج سے نوجوانوں کے تعلیم حاصل کئے بغیر چلے جانے کے ساتھ جڑی ہوئی ہے:

17) سولہ سے 18 سال کے نوجوان سال میں اوسطاً £370 ٹراسپورٹ پر خرچ کرتے ہیں ۔

18) سولہ سے 18 سال کے 47% نوجوان اس خرچ کے سلسلہ میں دقت محسوس کرتے ہیں ۔

19) 16 سے 24 سال کے 6% نوجوان ٹرانسپورٹ کے مسائل کی وجہ سے ٹریننگ یا فردرایجوکیشن کے مواقع سے انکار کر دیتے ہیں ۔

20) جو لوگ پبلک ٹرانسپورٹ پر انحصار کرتے ہیں، ان کے لئے ہسپتال پہنچنا خاص طور پر مشکل ہوتا ہے، اور اس کے نتیجہ میں ہو سکتا ہے کہ وہ صحت کے سلسلہ میں اپوائنٹمنٹوں پر نہ پہنچ سکیں:

21) کار کے بغیر 31% لوگوں کو، کار والے لوگوں کے 17% کے مقابلے میں اپنے مقامی ہسپتال تک سفر کرنے میں مشکلات پیش آتی ہیں ۔

22) کاروں کے بغیر 7% لوگ کہتے ہیں کہ ٹرانسپورٹ کے مسائل کی وجہ سے ان کو پچھلے 12 مہینے میں میڈیکل مدد سے محروم ہونا پڑا ہے، یا انہوں نے میڈیکل مدد لینے سے انکار کر دیا ہے یا نہ لینے کا فیصلہ کیا ہے ۔

23) یہ عمومی آبادی کی شرح سے دوگنا ہے ۔

24) پست ترین سماجی طبقہ سے آنے والے بچوں کے سڑک کے حادثات میں مرنے کا امکان اعلیٰ ترین سماجی طبقہ کے بچوں کے مقابلہ میں پانچ گنا زیادہ ہے ۔

25) کاروں کے بغیر 16% لوگ سپر مارکیٹوں تک پہنچنے میں، کاروں والے لوگوں کے 6% کے مقابلے میں، زیادہ مشکل محسوس کرتے ہیں ۔

26) ناقص ٹرانسپورٹ، دوسری سرگرمیوں کے ایک بڑے سلسلہ میں لوگوں کی شرکت پر بھی اثر انداز ہو سکتی ہے ۔

27) ان میں دوستوں اور خاندان سے ملنا، رضاکارانہ کام کرنا اور دیکھ بھال کا کام کرنا، مذہبی سرگرمیاں، ورزش اور ثقافتی سرگرمیاں شامل ہیں ۔

28) کار کے بغیر لوگوں میں سے 18% کو، کار مالکوں میں سے 8% کے مقابلہ میں، ٹرانسپورٹ کی وجہ سے دوستوں اور خاندان والوں سے ملنا زیادہ مشکل محسوس ہوتا ہے ۔

29) کار کے بغیر لوگوں کے، فرصت کے مشاغل کے مراکز (9%) اور لائبریریوں میں پہنچنے (7%) میں مشکل محسوس کرنے کا امکان بھی دوگنا زیادہ ہے ۔

30) تین گھرانوں میں سے تقریباً ایک گھرانہ کو کار تک دسترس حاصل نہیں ہے ۔

31) وہ ادھر ادھر جانے کے لئے زیادہ تر پیدل چلنے پر انحصار کرتے ہیں، اور بسوں، خاندان اور دوستوں سے لفٹ اور ٹیکسیوں پر بھی ۔

32) سائکلنگ اور ریل ان کے سفر کا بہت چھوٹا حصہ ہوتا ہے ۔

33) کام، تعلیم حاصل کرنے، صحت کی دیکھ بھال اوردوسری کلیدی سرگرمیوں تک پہنچنے میں لوگوں کو تین قسم کی رکاوٹیں پیش آسکتی ہیں:

34) رسائی اوردستیابی: لوگ کلیدی مقامات تک معقول وقت میں یقینی طورپر اورمحفوظ طریقے سے نہیں پہنچ سکتے۔

35) اس کی وجہ نیٹ ورک کا ناقص پھیلاؤ، پبلک ٹرانسپورٹ کا ناقص تواتر اوراس کا قابل اعتبارنہ ہونا یا قابل رسائی سہولتوں کا فقدان ہوسکتا ہے۔

36) بسوں میں سے صرف 20 فیصد اورٹرینوں میں سے صرف 10 فیصد، ڈس ابلٹی ڈس کری می نیشن ایکٹ کے تحت نئے ضوابط رسائی پرپورا اترتے ہیں۔

37) اس کے علاوہ، دیہی علاقوں میں کار کے بغیر رہنے والے لوگ، بس سٹاپوں تک پہنچنے کے لئے چلنے کے زیادہ لمبے فاصلوں، اورسروس کے کم تواتر کی وجہ سے خاص طورپر سنگین مسائل سے دوچارہوتے ہیں۔

38) لاگت: لوگ ذاتی یا پبلک ٹرانسپورٹ برداشت کرنے کی سکت نہیں رکھتے۔

39) بسوں کے کرائے پچھلے پندرہ سال میں تقریبا ایک تہائی بڑھ گئے ہیں۔

40) کم آمدنی والے گھرانے جن کے ہاں کار ہے بھی، اپنے گھرانے کے ہفتہ وارخرچ کا تقریبا ایک چوتھائی موٹرنگ پر خرچ کرتے ہیں۔

41) سفر کے افق: لوگ لمبی مدت یا لمبے فاصلے کے لئے سفر کرنے سے ہچکچاتے ہیں، یا شاید ٹرانسپورٹ کی سروسوں پر اعتماد نہیں رکھتے یا ان سے پوری طرح آگاہ نہیں ہیں۔

B.3. Baseline System Output

1) آپ حاصل کرسکتے ہیں ان کو اپنے سوشل سکیورٹی آفس سے ہے۔

2) ناقص ٹرانسپورٹ contributes کو سماجی اخراج میں دو طریقوں سے ہے۔

3) پہلے تو یہ restricts حاصل کرنے کی سرگرمیوں سے کہ enhance لوگوں کی زندگی chances ،، مثال کے طورپر کام کرتے ہیں، اور، صحت کی دیکھ بھال، کھانے کی اشیاء کی شاپنگ، اوردوسری کلیدی سرگرمیوں سے ہے۔

4) دوسری، محروم کمیونٹیوں میں مبتلا ہیں جنھیں سے استعمال سے اموات، pollutionاورisolation ہے جو اس کے نتیجہ می ں سے زندگی گزارنے کے مصروف سڑکوں پر ہیں۔

5) ایسی ہیں جن کی تعداد contributors کو سماجی اخراج سے متعلق ہے۔

6) ناقص ٹرانسپورٹ ہے تو ان میں سے ایک ہے۔

7) بہت سے لوگ experiencing سماجی اخراج نہیں کرے گا جس میں چالیس ناقص ٹرانسپورٹ ہے۔

8) تاہم، ناقص ٹرانسپورٹ ہوسکتا ہے پر میں restricting تک رسائی حاصل کرنے کا موقع ملتا ہے۔

9) یہ کرسکتے ہیں اس لئے undermine اہم حکومت objectives پر ویلفیئر کے ساتھ کام کرنے، raisingتعلیمی achievementاورحد صحت inequalities ہے، اوراخراجات کے لئے لوگوں کی، کاروباری اداروں، کمیونٹیوں اورلوگوں کو اسٹیٹ ہے۔

10) ٹرانسپورٹ بھی ہوسکتا ہے کہ کسی ایک اہم barrier کو accessing کام کرتے ہیں:

11) دو میں سے پانچ jobseekers کہتے ہیں کہ نہ ٹرانسپورٹ ہے barrier کو حاصل کرنے کا کام ہے۔

12) میں سے ایک چار jobseekers میں کہا گیا ہے کہ اس کی قیمت کے ٹرانسپورٹ ہے کسی مسئلہ پر میں آپ کو interviews ہے۔

13) میں سے ایک چارنوجوان لوگوں کو کے لئے درخواست نہیں دی تو ایک خاص طور پر کام میں گزشتہ بارہ ماہ کی وجہ سے ٹرانسپورٹ مسائل پیدا ہو سکتے ہیں ۔

14) میں سے ایک دس لوگوں میں لوانکم علاقوں کو دے پ کی ۔ میں اس کے آخری پر بارہ ماہ کی وجہ سے ٹرانسپورٹ ہے ۔

15) نوعمر افراد ڈرائیونگ لائسنس ہیں دو دفعہ کے طور پر اس کا حاصل کرنے کے کاموں سے ان لوگوں کے بغیر ہو سکتا ہے ۔

16) ناقص ٹرانسپورٹ ہے کو ٹھیکے سے نوعمر افراد dropping کے کالج کے :

17) sixteen سے 18 سال سے صرف ہفتے میں اوسطا £370 لگانے پر ٹرانسپورٹ ہے ۔

18) forty کی سات فی صد سے 16 سے 18 سال سے تجربہ میں اس سے آئی ۔

19) چھ فی صد سے 16 سے 24 سال سے یہ نتیجہ بھی ٹریننگ یا مزید تعلیم کے مواقع کی وجہ سے مسائل سے ٹرانسپورٹ ہے ۔

20) ان لوگوں کے لئے جو پر پبلک ٹرانسپورٹ، میں آپ کو ہسپتال ہے خاص طور پر مشکل اور مہیا کر سکتی ہے روانہ صحت appointments :

21) تاہم تیس سال کے مقابلے میں ایک فی صد لوگوں کے بغیر کسی گاڑی کو مشکلات درپیش ہیں ان کے راستے کو ان کے مقامی ہسپتال، ٹھیک کرنے کے لئے 17 فی صد لوگوں کی ایک گاڑی ہے ۔

22) سات فی صد لوگوں کے بغیر گاڑیاں کہتے ہیں کہ وہ نہیں ہوں گے، دے پ کی، یا ان کے بس کی بات نہیں ہے کہ وہ طبی مدد میں گزشتہ بارہ ماہ کی وجہ سے ٹرانسپورٹ مسائل پیدا ہو سکتے ہیں ۔

23) یہ بات ہے double ریٹ میں عام ہوتا ہے ۔

24) بیچوں کے کم سے کم سے کم سوشل کلاس ہیں ان میں سے پانچ مرتبہ اس کے نتیجے میں سڑکوں پر ہونے والے حادثات سے کم ہوتا ہے کہ وہ اس بات کا ہے جس کی وجہ سے مقابلے سوشل کلاس ہے ۔

25) sixteen فی صد لوگوں کے بغیر گاڑیاں تلاش کرنے کی supermarkets مشکل ہوتا ہے، ٹھیک سے چھ فی صد لوگوں کے ساتھ ساتھ لیتے جائیے ۔

26) ناقص ٹرانسپورٹ بھی کر سکتے ہیں اس بات پر اثر پڑتا ہے کہ لوگوں کے داروں کی شمولیت میں اس طرح کے دیگر سرگرمیوں سے ہے ۔

27) جن میں ملاقات کے دوستوں اور خاندان، volunteering کے بارے میں اور، مذہبی سرگرمیوں، ورزش اور ثقافتی سرگرمیوں کا ہے ۔

28) eighteen فی صد لوگوں کے بغیر کسی گاڑی کو پتہ چلتا ہے کہ ملاقات کے دوستوں اور خاندان مشکل کی وجہ سے ٹرانسپورٹ، ٹھیک سے آٹھ فی صد کے لئے گاڑی owners ہے ۔

29) لوگوں کے بغیر گاڑیاں کے علاوہ بھی دو دفعہ کی حیثیت سے کام کرنے کا امکان مشکل میں آپ کو تفریحی مراکز) 9 فی صد (اور لائبریریاں) کے لئے سات فی صد (۔

30) تقریبا میں سے ایک تین گھرانوں کی ضرورت نہیں ہوتی حاصل کرنے کی ایک گاڑی ہے ۔

31) ان کا primarily پر چلنے کی حاصل کرنے کے لئے ہیں، لیکن اس کے ٹرینوں پر، lifts سے اپنے خاندان والوں اور دوستوں اور taxis ہے ۔

32) سائیکلنگ اور rail اس پر عمل ہوتا ہے جو چھوٹے fraction کے journeys ہے ۔

33) جا سکتی ہے ۔ لوگوں کے لئے بہت کچھ کر سکتے ہیں جن سے تین طرح کے barriers کو accessing کام سیکھنے، صحت کی دیکھ بھال اور دوسری کلیدی سرگرمیوں کا :

34) تک رسائی اور: availability اس سے لوگوں کو حاصل نہیں کر سکتے کہ اہم مقامات پر معقول مدت میں، reliably اور کو

-

35) یہ ہو سکتا ہے اوراس کو ناقص نیٹ ورک کے احاطہ، frequency، اور reliability کے پبلک ٹرانسپورٹ یا نہ قابل رسائی میں مدد کے لئے ہے۔

36) صرف 20 فی صد سے ٹرینوں اوردس فی صد سے کم کربسوں کو پورا کرنے کے لئے نئے accessibility قوانین کے تحت اس معذوری کے تعصب اختیارحاصل ہے۔

37) اس کے علاوہ لوگوں میں رہنے والے دیہاتی علاقوں میں بغیرکسی گاڑی سے خاص طورپر جسم کے مسائل کو زیادہ چلنے کی distancesکو بس سے ملنا بند ہوگیا ہے، اورکم سروس frequency ہے۔

38) قیمت: اس سے لوگوں کو نہیں کر سکتے حیثیت ذاتی یا نقلوحمل کے عوامی ذرائع ہیں۔

39) بس کے اخراجات مثلاً کرائے نے risen کی طرف سے تقریبا ایک تہائی کمی کرنا میں آخری fifteen سال ہے۔

40) لوانکم گھرانوں کا کارامدادی کام کے تقریبا ایک چوتھائی کمی کرنا ان کے لئے ہر ہفتے گھر expenditure پر motoring ہے۔

41) british horizons: لوگوں کی ضرورت ہے unwilling کو british طویل journey اوقات یا distances یا ہو سکتا ہے کہ نہ ٹرسٹ میں، یا انسیت میں ٹرانسپورٹ کے لئے تجویزکرتی ہیں۔

B.4. Distance-Based System Output

1) آپ ان سے حاصل کر سکتے ہیں اپنے سوشل سکیورٹی آفس کے۔

2) ناقص ٹرانسپورٹ contributes سماجی اخراج کو دو طریقوں سے میں ہے۔

3) اس سے پہلے restricts سرگرمیوں تک رسائی ہو کہ enhance لوگوں کی زندگی chances، جیسے کام سیکھنے،، صحت کی دیکھ بھال کھانے کی اشیاء، شاپنگ، اوردوسری کلیدی سرگرمیوں کا۔

4) دوسری، محروم کمیونٹیوں کے چالیس سے اموات واقع ہوتی ہیں جنھیں استعمال کریں، pollution اور isolation کر سکتے ہیں جس سے اس کے نتیجہ میں رہنے والے مصروف سڑکوں کے نزدیک ہے۔

5) ایسی ہیں جن کے نمبر contributors سماجی اخراج کو۔

6) ناقص ٹرانسپورٹ محض ہے ان میں سے ایک ہے۔

7) بہت سے لوگ experiencing سماجی اخراج چالیس ناقص ٹرانسپورٹ نہیں کرے گا۔

8) تاہم، ناقص ٹرانسپورٹ ہو سکتا ہے پر restricting میں موقع تک رسائی ہو۔

9) یہ کر سکتے ہیں اس لئے کلیدی حکومت undermine objectives ویلفیئر کے ساتھ کام کرنے کے بارے میں، تعلیمی اور raising achievement حد inequalities ہے، اورصحت کے لئے اخراجات میں سے افراد، کاروباری اداروں، اور جماعتوں کے لوگوں کو اسٹیٹ ہے۔

10) ٹرانسپورٹ بھی ہو سکتا ہے کہ کسی ایک کو نمایاں barrier accessing کام کرتے ہیں:

11) دو میں سے پانچ jobseekers کہتے ہیں کہ نہ ٹرانسپورٹ barrier مل رہا ہے کہ ان کا کام ہے۔

12) چارمیں سے ایک jobseekers میں کہا گیا ہے کہ ٹرانسپورٹ کی قیمت میں کمی کرنے کے لئے ایک مسئلہ interviews کو مل رہا ہے۔

13) چارمیں سے ایک نوجوان لوگوں کو ایک خاص طورپر کام کے لئے درخواست نہیں دی تو گزشتہ بارہ ماہ میں ٹرانسپورٹ کی وجہ سے مسائل پیدا ہو سکتے ہیں۔

14) 10 میں لوگوں میں سے ایک لوانکم علاقوں دے پ کی ہے۔ گزشتہ پر بارہ ماہ میں ٹرانسپورٹ کی وجہ سے ہے۔

15) نوعمر افراد ہیں ڈرائیونگ لائسنس حاصل کرنے کا امکان کے طورپر دو دفعہ کاموں کے بغیر ان لوگوں سے ہے۔

16) ناقص ٹرانسپورٹ ہے نوعمر افراد کو ٹیکے dropping کالج کے :

17) sixteen کی 18 سال سے امدادی کام ہفتے میں اوسطا £370 ایک سال کی ٹرانسپورٹ پر ہے ۔

18) forty کی سات فی صد سے 16 سے 18 سال سے میں اس تجربہ کے ساتھ آئی ۔

19) چھ فی صد سے 16 سے 24 سال سے زیادہ ہے اس کے ٹریننگ یا فردایجوکیشن کے مواقع کی وجہ سے ٹرانسپورٹ کے ساتھ مسائل کی ہے ۔

20) ان لوگوں کے لئے پر ہے جو پبلک ٹرانسپورٹ ، ہسپتالوں میں آپ کو خاص طور پر مشکل ہے ، اورصحت appointments روزگار مہیا کر سکتی ہے :

21) تاہم تیس سال کی ایک فی صد لوگوں کی ایک گاڑی نے سفر کے بغیر مشکلات درپیش ہیں ان کو ان کے مقامی ہسپتال ، کو ٹھیک 17 فی صد لوگوں کی ایک گاڑی کے ساتھ ۔

22) کے لئے سات فی صد لوگوں کی گاڑیاں بغیر کہتے ہیں کہ وہ نہیں ہوں گے ، دے پ کی یا نہیں ، میں آج طبی مدد دیتا ہے کہ وہ گزشتہ بارہ ماہ میں ٹرانسپورٹ کی وجہ سے مسائل پیدا ہو سکتے ہیں ۔

23) double ہے اس میں ریٹ پر جنرل ہوتا ہے ۔

24) بیچوں کے زیریں سے سوشل کلاس ہیں زیادہ امکان اس بات کا پانچ مرتبہ اس کے نتیجے میں ان لوگوں سے سڑکوں پر ہونے والے حادثات میں کیلئے اعلٰی سوشل کلاس سے ہے ۔

25) sixteen فی صد لوگوں کی گاڑیاں supermarkets تک رسائی کے بغیر تلاش کرنا مشکل ہوتا ہے ، کے ساتھ ٹھیک چھ فی صد لوگوں کی گاڑیاں کے ساتھ ۔

26) ناقص ٹرانسپورٹ بھی کر سکتے ہیں اس بات پر اثر پڑتا ہے کہ لوگوں کے داروں کی شمولیت ایک میں اس طرح کے دیگر سرگرمیوں کا ۔

27) جن میں ملاقات کے volunteering ، دوستوں اور خاندان کے بارے میں اور سرگرمیوں ، مذہبی ،ثقافتی سرگرمیوں اور ورزش ہیں ۔

28) eighteen فی صد لوگوں کی تلاش بغیر کسی گاڑی کے ساتھ ملاقات کے دوستوں اور خاندان کی وجہ سے مشکل ٹرانسپورٹ ، کے ساتھ ٹھیک آٹھ فی صد کے لئے owners گاڑی ہے ۔

29) لوگ گاڑیاں کے بغیر بھی ہوتے ہیں جیسے جیسے دو دفعہ ہونے کا امکان مشکل تفریحی مراکز میں آپ کو) 9 فی صد اور لائبریریاں ( ) کے لئے سات فی صد (۔

30) تقریبا تین گھرانوں میں سے ایک کی ضرورت نہیں ہوتی کسی گاڑی تک رسائی ہو ۔

31) وہ primarily پیدل چلنے پر بھروسہ حاصل کرنے کے بھی ہیں ، لیکن ٹرینوں پر ،lifts آپکی فیملی اور دوستوں اور taxis ہے ۔

32) سائیکلنگ اور rail آن ایک ہے جو چھوٹے fraction journeys ان کے لئے ممکن ہے ۔

33) 9 ۔ لوگوں کا سامنا ہو کر سکتے ہیں تین طرح کے barriers accessing کو کام سیکھنے ،، صحت کی دیکھ بھال اور دوسری کلیدی سرگرمیوں کا :

34) تک رسائی حاصل نہیں کر سکتے اور availability : اس سے لوگوں کو معقول مدت میں اہم مقامات ، اور reliably کو ۔

35) ہو سکتا ہے کہ اس کو کہا کے احاطہ ناقص نیٹ ورک ،frequency ، اور پبلک ٹرانسپورٹ کے reliability یا نہ سہولتیں قابل رسائی ہے ۔

36) صرف 20 فی صد سے کم دس فی صد سے ٹرینوں اور بسوں کو پورا کرنے کے لئے accessibility نئے قوانین کے تحت تعصب ایکٹ میں مشکلات پیش آتی ہیں ۔

37) اس کے علاوہ دیہاتی علاقوں میں رہنے والے لوگوں کے بغیر خاص طور پر مسائل کا سامنا ہو کسی گاڑی میں جسم کو پیدل چلنے کے بعد distances کو بس سے ملنا بند ہو گیا ہے ، اور کم frequency سروس ہے ۔

38) قیمت : اس سے لوگوں کو ذاتی حیثیت یا پبلک ٹرانسپورٹ نہیں کر سکتے ہیں ۔

39) بس کے اخراجات مثلا کرائے کی طرف سے risen نے تقریبا ایک تہائی کمی کرنا میں fifteen گزشتہ سالوں سے چل رہے ہیں ۔

40) لو انکم گھرانوں کا امدادی گاڑی کے استعمال سے تقریبا ایک چوتھائی کمی کرنا ان کے خاندان کے ہفتہ expenditure motoring پر ہے ۔

41) british horizons : لوگوں کی ضرورت ہے british unwilling کو طویل journey یا اوقات distances یا ہو سکتا ہے کہ میں نہ ٹرسٹ ، یا انسیت ، ٹرانسپورٹ کے ساتھ ہیں ۔

B.5. Transformation System Output

1) آپ کو اپنے سوشل سکیورٹی آفس سے ہی ان سے حاصل کر سکتے ہیں ۔

2) ناقص ٹرانسپورٹ دو طریقوں سے سماجی اخراج کو contributes ۔

3) سب سے پہلے اس کی سرگرمیوں کا مطلب یہ ہوا کہ امکانات زندگی لوگوں کے enhance ، مثال کے طور پر کام سیکھنے ، صحت کی دیکھ بھال ، کھانے کی اشیاء کی شاپنگ ، اور دوسری کلیدی سرگرمیوں تک رسائی restricts ۔

4) دوسری ، محروم کمیونٹیوں ہیں جنھیں غیر متناسب حد تک pedestrian اموات واقع ہوتی ہیں ، pollution اور isolation جو مصروف سڑکوں کے پاس ایسا سینٹر زندگی سے ہو سکتا ہے سے بہت کم ہے ۔

5) اس کی وجہ سے سماجی اخراج کو contributors کا نمبر ہے ۔

6) ناقص ٹرانسپورٹ ان میں سے صرف ایک ہے ۔

7) بہت سے لوگ سماجی اخراج experiencing ناقص ٹرانسپورٹ سے بہت کم اثر نہیں پڑے گا ۔

8) تاہم ، ہو سکتا ہے کہ ناقص ٹرانسپورٹ موقعہ تک رسائی restricting میں اہم بنیاد پر کیا جا سکتا ہے ۔

9) اس لئے تعلیمی achievement اور حد تک صحت inequalities بلند کرے گی ، کام کرنے کی بہبود کے بارے میں اہم حکومت objectives undermine بھی ہو سکتی ہے ، اور انفرادی ، کاروباری اداروں ، جماعتوں اور لوگوں کو اسٹیٹ کے لئے اخراجات میں مدد ملی ہے ۔

10) ٹرانسپورٹ کام accessing کا نمایاں barrier جا سکتی ہیں :

11) دس میں سے پانچ jobseekers دو کہتے ہیں کہ ٹرانسپورٹ کے نہ ملازمت حاصل کرنے کے بارے میں ایک barrier ہے ۔

12) چار jobseekers میں سے ایک کا کہنا ہے کہ ٹرانسپورٹ کا خرچہ پورا کرنے کے لئے ایک مسئلہ interviews کو مل رہا ہے ۔

13) چار نوجوانوں میں سے ایک کی کوئی خاص کام کی وجہ سے ٹرانسپورٹ کے مسائل ہیں اور گزشتہ بارہ ماہ میں کے لئے درخواست نہیں ہوتی ہے ۔

14) کم تنخواہ والے بحران کے علاقوں میں دس افراد میں سے ایک کام کی وجہ سے ٹرانسپورٹ گزشتہ پر بارہ ماہ کے دوران مسترد ہے ۔

15) ڈرائیونگ لائسنس کے ساتھ نوجوان لوگ اچھی طرح جیسے کاموں کے لئے ان سے بغیر حاصل ہوتی ہیں ۔

16) ناقص ٹرانسپورٹ کے نوجوانوں کو کی گئی ایک تحقیق کے بعد کالج کے dropping لنکڈ ہے :

17) sixteen 18 ۔ کی سال کی عمر میں اوسطا پر £370 ایک سال کی ٹرانسپورٹ میں صرف کرتے ہیں ۔

18) عمر 16 سال کی فی صد forty کی سات 18 ۔ کی سال کی عمر کا تجربہ دشواری کے لئے اس کی قیمت میں مبتلا ہیں ۔

19) چھ سال کی فی صد سے 24 سال سے روکنے کی وجہ سے ٹرانسپورٹ کے ساتھ مسائل کی جانب تعلیم ٹریننگ یا مزید کرنے کا ارادہ رکھتی ہے ۔

20) جو لوگ کو چاہیئے کہ وہ ایسی پبلک ٹرانسپورٹ پر ، ہسپتالوں کو مل رہا خاص طور پر مشکل ہے ، اورروانہ نہ صحت کی ملاقاتوں لاحق ہو سکتے ہیں : کے بارے میں

21) ایک کار کے بغیر لوگوں کے فی صد thirty پر مبنی مشکلات بھی ان کے مقامی ہسپتال کے لئے کام کر رہی ہے ، ٹھیک سے ایک کار کے ساتھ ہر قسم کے لوگ فی صد 17 کرنے کے راستے سفر کرتے ہیں ۔

22) گاڑیوں کے بغیر لوگوں کے فی صد سے کم سات کہتے ہیں کہ وہ وقت نہ ہونے کی وجہ سے ٹرانسپورٹ کے مسائل ہیں اور گزشتہ بارہ ماہ سے زائد طبی مدد طلب کرنے یا ، مسز بڑھنے ، روانہ نہ ہو گیا ہے ۔

23) اس میں عام آبادی میں شرح ہے double ۔

24) اس کے زیریں سماجی طبقہ بیچوں کو پانچ مرتبہ زیادہ امکان اس بات کا ہے کہ سڑکوں پر ہونے والے حادثات میں کے مقابلے میں سماجی طبقہ ان لوگوں سے مرنے والے میں کام کر رہے ہیں ۔

25) گاڑیوں کے بغیر لوگوں کے فی صد sixteen سپر مارکیٹ تک رسائی ، مشکل ہو جاتا ہے کے لئے ٹھیک سے گاڑیوں کے ساتھ ہر قسم کے لوگ فی صد سے کم چھ میں مبتلا ہیں ۔

26) ناقص ٹرانسپورٹ کے علاوہ دیگر سرگرمیوں کے بارے میں لوگوں کے داروں کی شمولیت کو بھی متاثر کر سکتی ہے ۔

27) اور volunteering دیکھ بھال کر رہے ہیں ، جس میں دوستوں اورخاندان seeing مذہبی سرگرمیوں ، ورزش اور ثقافتی سرگرمیوں کا انتظام کیا جا سکتا ہے ۔

28) ایک کار کے بغیر لوگوں کے فی صد eighteen تلاش seeing دوستوں اورخاندان کی وجہ سے ٹرانسپورٹ کرنا مشکل ہے ، ٹھیک سے کار ذاتی کے لئے فی صد سے کم آٹھ میں مبتلا ہیں ۔

29) گاڑیوں کے بغیر لوگ بھی اچھی طرح سے یہ تاثر پیدا ہوتا ہے کہ اس کا امکان بہت کم ہے آپ کے لئے یہ مشکل تفریحی مراکز) فی صد 9 (اور لائبریریاں) فی صد سے کم سات (کو مل رہا ہے ۔

30) تین گھرانوں میں تقریبا ایک ایک گاڑی ہے تو نہیں کرتا ۔

31) وہ taxis اوراپنے خاندان والوں اوردوستوں سے lifts ، ٹرینوں میں حکومت کے بارے میں بھی ملے لیکن وہ primarily پیدل مل بھی گردش کرنے کے بارے میں بات مد نظر رکھی ۔

32) cyclingاور rail نے اپنی journeys کا ایک مختصر fraction کرتے ہیں ۔

33) 9 ۔ ، صحت کی دیکھ بھال اوردوسری کلیدی سرگرمیوں : لوگ accessing کام سیکھنے کے لئے barriers کی تین قسمیں کا سامنا ہو سکتا ہے

34) رسائی اوردستیاب ہونا ۔ : لوگوں کو معقول مدت میں اہم علاقوں کو تو reliably اور کے ساتھ حاصل نہیں کر سکتے ہیں ۔

35) یہ ناقص نیٹ ورک کے احاطہ ، frequency، اورپبلک ٹرانسپورٹ کے reliability یا رسائی کی سہولتیں مہیا کی جانے والی نہ کرنے کے دوران کیا جا سکتا ہے ۔

36) ٹرینوں میں حکومت کی فی صد صرف 20 اوربسوں کی فی صد 10 کو معذوری کے تعصب ایکٹ کے تحت نئے accessibilityان قوانین کے الگ الگ ہوتی ہے ۔

37) اس کے علاوہ لوگوں میں rural کے اندرونی علاقوں میں ایک کار کے بغیر زندگی گزارنے والے کے لئے کام کر رہی ہے جس کے بعد پیدل distances بس سے ملنا بند ہو گیا ، اورکم تنخواہ سروس frequency خاص طور پر ، مسائل کا سامنا ہو ۔

38) قیمت (: ذاتی یا پبلک ٹرانسپورٹ میں قوت برداشت نہیں کر سکتے ہیں ۔

39) بس اخراجات مثلا کرائے گزشتہ fifteen سالوں میں تقریبا ایک تہائی کمی ہو گیا ہے ۔

40) کم تنخواہ والے آمدنی گھرانوں کے کہ کیا یہ بات سامنے آئی ہے کہ کار motoring پر اپنے ہفتہ وار گھریلو expenditure کے تقریبا ایک چوتھائی صرف کرتے ہیں ۔

41) سفر : horizons لوگ طویل journey اوقات یا distances کے سفر کو ہاتھ نہ ڈالنا چاہتے ہوں ہیں، یا کام کے بارے میں معلومات فراہم کریں گے ، یا familiarity سے ٹرانسپورٹ سروسز ٹرسٹ نہ ہو سکتا ہے ۔

Ata N., Jawaid B. and Kamran A. Rule Based English to Urdu Machine Translation [Conference]. - [s.l.] : In Proceedings of Conference on Language and Technology (CLT'07), Pakistan, 2007.

Baker P. [et al.] EMILLE, A 67-Million Word Corpus of Indic Languages: Data Collection, Mark-up and Harmonisation [Conference]. - [s.l.] : In Proceedings of the 3rd Language Resources and Evaluation Conference, pp. 819-825, LREC' 2002.

Bojar O., Straňák P. and Zeman D. English-Hindi Translation in 21 Days [Conference]. - [s.l.] : In Proceedings of ICON NLP Tools Contest, Pune, India, 2008.

Brown P. F. [et al.] A Statistical Approach to Machine Translation. [Conference]. - [s.l.] : Computational Linguistics, Vol. 16, No. 2, pp. 79–85, June, 1990.

Brown P. F. [et al.] The mathematics of statistical machine translation: Parameter estimation [Conference]. - [s.l.] : Computational Linguistics 19 pp. 263–311, 1993.

Chen S. and Goodman J. An Empirical Study of Smoothing Techniques for Language Modeling [Conference]. - [s.l.] : Harvard University, 1998.

Chen S. F. and Goodman J. An Empirical Study of Smoothing Techniques for Language Modeling [Conference]. - [s.l.] : Computer Speech and Language, 4(13):359-393, 1999.

Chiang D. A hierarchical phrase-based model for statistical machine translation [Conference]. - [s.l.] : In Proceedings of 43rd Annual meeting of the Association for Computational Linguistics (ACL), pp. 263–270, 2005.

Čmejrek M., Curín J. and Havelka J. Czech-English Dependency-based Machine Translation [Conference]. - [s.l.] : In Proceedings of the 10th Conference of The European Chapter of the Association for Computational Linguistics (ACL), pp. 83-90, Budapest, Hungary, 2003.

Collins M. Head-Driven Statistical Models for Natural Language Parsing [Report] : Ph.D. thesis / University of Pennsylvania. - 1999.

Eisner J. Learning non-isomorphic tree mappings for machine translation [Conference]. - [s.l.] : In Proceedings of 41st Annual Meeting of Association for Computational Linguistics (ACL), (companion volume, pp. 205–208), 2003.

Federico M. Machine Translation Basic Concepts [Conference]. - [s.l.] : Course Material for Statistical Machine Translation. Free University of Bozen. Bolzano, 2009.

Federico M., Bertoldi N. and Cettolo M. IRSTLM: an Open Source Toolkit for Handling Large Scale Language Models [Conference]. - [s.l.] : In Proceedings of Interspeech, Brisbane, Australia, 2008.

Jurafsky D. and Martin J. H. Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition [Conference]. - [s.l.] : Prentice-Hall:2000. ISBN 0-13-095069-6, 2000.

Khalilov M. and Fonollosa J. A. R. N-Gram-Based Statistical Machine Translation versus Syntax Augmented Machine Translation: Comparison and System Combination [Conference]. - [s.l.] : In Proc. of European Chapter of the ACL (EACL), pp. 424-432, Athens, Greece, 2009.

Knight K. and Koehn P. What's New in Statistical Machine Translation [Conference]. - [s.l.] : Tutorial at HLT/NAACL, 2004.

Koehn P. [et al.] Moses: Open Source Toolkit for Statistical Machine Translation [Conference]. - [s.l.] : In Proceedings of 45th Annual Meeting of the Association for Computational Linguistics, Demo and Poster Sessions, pp. 177-180, Prague, Czech Republic, 2007.

Koehn P. and Hoang H. Factored translation models [Conference]. - [s.l.] : In Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL), pp. 868–876, 2007.

Koehn P., Och F. J. and Marcu D. Statistical phrase-based translation [Conference]. - [s.l.] : In Proceedings of the Joint Conference on Human Language Technologies and the Annual Meeting of the North American Chapter of the Association of Computational Linguistics (HLT-NAACL), pp. 127–133, 2003.

Lopez A. Statistical Machine Translation [Conference]. - [s.l.] : In ACM Computing Surveys 40(3),: Article 8, pp. 1–49, 2008.

Manning C. D. and Schütze H. Foundations of Statistical Natural Language Processing [Conference]. - [s.l.] : MIT Press: 1999. ISBN 0-262-13360-1, 1999.

Marcu D. and Wong W. A phrase-based, joint probability model for statistical machine translation [Conference]. - [s.l.] : In Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 133–139, 2002.

Marcus M. P. [et al.] Treebank-3 [Conference]. - [s.l.] : Philadelphia, PA: Linguistic Data Consortium, University of Pennsylvania, 1999.

Och F. J. and Ney H. A comparison of alignment models for statistical machine translation [Conference]. - [s.l.] : In COLING '00: The 18th Int. Conf. on Computational Linguistics, pp. 1086–1090, Saarbrücken, Germany, August, 2000.

Och F. J. and Ney H. Improved Statistical Alignment Models [Conference]. - [s.l.] : In Proceedings of the 38th Annual Meeting of The Association for Computational Linguistics (ACL), pp. 440–447, 2000.

Och F. J. and Ney H. The alignment template approach to statistical machine translation [Conference]. - [s.l.] : Computational Linguistics, 30(4), pp.417–449, 2004.

Och F. J. Minimum Error Rate Training for Statistical Machine Translation [Conference]. - [s.l.] : In Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics (ACL), Japan, Sapporo, 2003.

Och F. J., Tillmann C. and Ney H. Improved alignment models for statistical machine translation [Conference]. - [s.l.] : In: Proc. of the Joint SIGDAT Conf. on Empirical Methods in Natural Language Processing and Very Large Corpora, pp. 20–28, University of Maryland, College Park, MD, 1999.

Papineni K. [et al.] BLEU: a method for automatic evaluation of machine translation [Conference]. - [s.l.] : In Proceedings of 40th Annual meeting of the Association for Computational Linguistics (ACL), pp. 311–318, 2002.

Ramanathan A. [et al.] Simple Syntactic and Morphological Processing Can Help English-Hindi Statistical Machine Translation [Conference]. - [s.l.] : International Joint Conference on NLP (IJCNLP08), Hyderabad, India, 2008.

Shannon C. E. A Mathematical Theory of Communication [Conference]. - [s.l.] : In (Vol. vol. 27, pp. 379-423 and 623-656). University of Pennsylvania, 1948.

Stolcke A. SRILM - An Extensible Language Modeling Toolkit [Conference]. - [s.l.] : In Proceedings of the International Conference on Spoken Language Processing, volume 2, pp. 901–904, 2002.

Toutanova K. and Manning C. D. Enriching the Knowledge Sources Used in a Maximum Entropy Part-of-Speech Tagger [Conference]. - [s.l.] : In Proceedings of the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora (EMNLP/VLC), pp. 63-70, 2000.

Vogel S., Ney H. and Tillmann C. HMM-based word alignment in statistical translation [Conference]. - [s.l.] : In: COLING '96: The 16th Int. Conf. on Computational Linguistics, pp. 836–841, Copenhagen, Denmark, 1996.

Yamada K. and Knight K. A syntax-based statistical translation model [Conference] // 39th Annual meeting of the Association for Computational Linguistics (ACL). pp. 523-530. - 2001.

Zens R. Kontextabhängige Statistische Ü bersetzungsmodelle [Conference]. - [s.l.] : Diploma thesis,Computer Science Department, RWTH Aachen, Aachen, Germany, June, 2008.

Zens R., Och F. J. and Ney H. Phrase-Based Statistical Machine Translation [Conference]. - [s.l.] : Proc. M. Jarke, J. Koehler, G. Lakemeyer, editors, 25th German Conf. on Artificial Intelligence (KI2002), Vol. 2479 of Lecture Notes in Artificial Intelligence (LNAI), pp. 18–32, Aachen, Germany, September 2002. Springer Verlag., 2002.

Zollmann A. [et al.] Systematic comparison of Phrase-based, Hierarchical and Syntax-Augmented Statistical Machine Translation [Conference]. - [s.l.] : In: COLING '08: The 22nd Int. Conf. on Computational Linguistics, pp. 1145–1152, Manchester, UK, 2008.