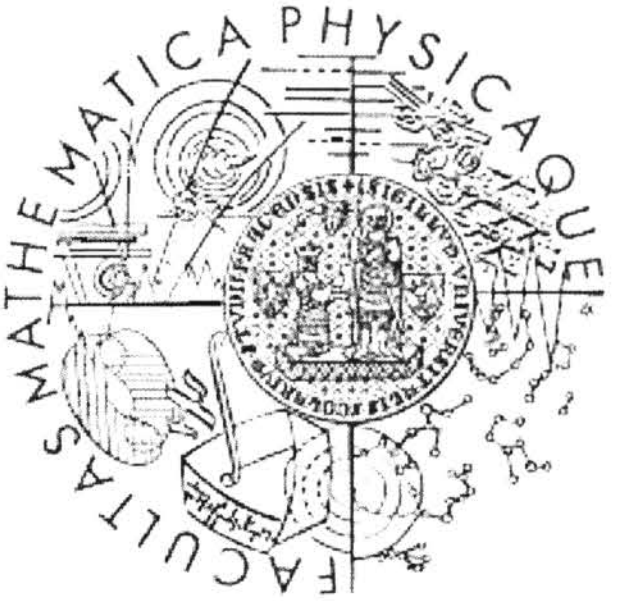
Charles University in Prague

Faculty of Mathematics and Physics

DIPLOMA THESIS



Naila Ata

Analyzing Errors and Chances to improve English to Urdu Phrase-based Translation

Institute of Formal and Applied Linguistics

Supervisor: RNDr. Dr. Ondřej Bojar
Study Programme: Computer Science
Study Field: Mathematical Linguistics

Prague, 2010

I certify that this diploma thesis is all my own work, and that I used only the cited literature. The thesis is freely available for all who can use it.

Prague, August 24, 2010

Naila Ata

Contents

Introduction 8

1.1 Goal of the thesis: 9

1.2 Structure of the thesis: 9

Statistical Machine Translation 11

2.1 Statistical Modeling of translation System: 11

Translation Modeling Overview: 12

2.2 Word-based Translation Modeling: 13

Definitions: 13

Translation Model Probability Estimation: 14

2.3 Phrase-Based Translation Modeling: 15

A Simple Phrase based Model: 15

Log-linear Model 16

Log-Linear, Phrase-Based Translation Models: 17

Factored Translation Model 19

3.1 Introduction: 19

Factors: 19

3.2 Translation in Factored Model: 20

3.3 Statistical Model of Factored Translation: 21

Training: 21

Combination of Component 21

Decoding: 22

Further enhancements: 22

Use of Automatic word classes: 22

Integrated Recasing: 22

MT Error Evaluation 23

4.1. Related Work 23

4.2. Annotation Scheme: 24

Explanation: 24

4.3. Examples 27

4.4. Analysis of Annotation....28
Data Acquisition....30
5.1 English-Urdu Parallel Data Acquisition:....30
Experimental Framework....32
6.1 The Corpora....32
Preprocessing of Corpus:....32
Factored Corpus:....33
6.2 Software tools:....33
6.3 Tuning and Evaluation....34
Translation Evaluation....34
Tuning....34
Searching for better Phrase based Machine Translation....35
7.1 Reasons of Errors:....35
7.2 Ways to Improve translation:....35
Modeling target side syntactic structure:....36
Urdu Tagging:....36
Case Markers:....37
Reordering:....37
Related Work:....38
Reordering Model in Urdu....38
Dependency Parsing:....39
Results....41
8.1 Experiments With Different Corpus:....41
8.2 Experiments with Factored translation:....44
Impact of Generalized tagging for Urdu:....45
Experiment with Reordering:....46
Conclusion and Future Dimension....47
Future Dimensions....47
Bibliography....48

Title : Analyzing Errors and Chances of Improving English to Urdu Phrase-Based Translation
Author: Naila Ata
Department: Institute of Formal and Applied Linguistics
Supervisor: RNDr. Ondřej Bojar, Ph.D
Supervisor's email address: bojar@ufal.mff.cuni.cz

Abstract:
The aim of the thesis is to analyze errors in English to Urdu phrase-based or hierarchical phrase-based machine translation, and to propose and evaluate a few possible improvements in translation quality.
The first step consists of setting up and running a suitable MT system, e.g. Moses or Joshua, including the necessary collection of a small training and evaluation parallel corpus. A thorough manual analysis of the system output of the given test corpus should indicate the most severe problems of the translation quality. The thesis should then attempt to tackle the identified issues by e.g.: (1) pre-processing of input English, such as word reordering, (2) preprocessing the training corpus in order to reduce unnecessary lexical ambiguity, (3) using additional factors (in Moses factored translation) to better model target-side morphological coherence. For any of the options, either rule-based or statistical approaches may be applied. The utility of the proposed modifications to the translation pipeline have to be evaluated by both automatic MT metrics as well as human judgments on a small subset of the test corpus.

Chapter 1

Introduction

In the current information age, with the availability of huge electronic data, machine translation (MT) has gained more importance and has become one of the most active research areas in computational linguistics.

Traditionally, machine translation systems are divided into two broad classes: Rule based machine translation system and Statistical machine translation system.

Rule-based MT is based on linguistic knowledge and it translates sentence from source language to target language using linguistic information (morphology, grammar, dictionaries etc). Translation is breakdown into three steps: analysis, analysis of source language sentence; transfer, intermediate representation of sentence; and synthesis, generation of target language sentence.

Statistical machine translation (SMT) is a paradigm for translating text from one language to another, based on statistical methods. Statistical models are automatically estimated from the parallel corpus. A statistical machine translation system can be divided into three parts: a translation model, which defines the correspondences between words and phrases of source language and target language; a language model which describes the degree of fluency of a sentence in target language; and a decoder, which tries to find best target language sentence by decoding source language sentence with the help of translation model and language model.

There are pros and cons of both approaches. Rule-based systems are built on linguistic theories and are able to give result but it needs a lot of effort to engineer required linguistic information. For instance, syntactic rule-based MT requires the transformation rules to convert source language grammatical structures into target language grammatical structure, on contrary, SMT does not rely on any language specific details but it fails to capture complicated syntactic structure, especially when there are various differences between source language and target language.

Recently, there are attempts to incorporate syntactic information into Statistical MT to get better results. Especially, techniques for automatic extraction of grammar from the corpus and shallow parsing for word alignment improved the quality of generated sentence.

Phrase based systems are one of the best performing MT systems in statistical translation systems. Instead of calculating word translation probability, it learns translation probabilities of phrases of source language sentence to target language sentence.

1.1 Goal of the thesis:

In this thesis, we describe ways to improve phrase based machine translation from English to Urdu. Urdu is a Subject-Object-Verb SOV language while English is Subject-Verb-Object SVO language. Moreover, Urdu is morphologically richer language than English. This thesis attempts to describe errors related to linguistic differences between two langauges and discusses implemented ways to improve the translation.

Goal of this thesis includes:

To build a phrase based translation model for English to Urdu.

Analyze the error by manually annotating 200 test sentences, generated by base-line system

To identify most frequent errors and its source

Design and implement ways to remove frequently occurring errors.

1.2 Structure of the thesis:

- Chapter 2 gives an overview of statistical machine translation. It describes the basic types of translation models, particularly the current state-of-the-art phrase based translation model and log-linear translation model. These are the models that we use as our experimental framework.
- Chapter 3 explains the factored machine translation, and describes usage of various factors, specifically part-of-speech tag, lemma, and stem of a word.
- Chapter 4 describes the annotation scheme which is used to mark different types of errors. It starts with the introduction of Vilar [et.el., 2006] error annotation framework and then it explains the modifications done in the scheme and reasons of these modifications.
- Chapter 5 explains data issues related to English-Urdu parallel corpus. It also explains morphological variations between two languages and preprocessing ways to improve sentence alignment.
- Chapter 6 presents experimental framework to evaluate base-line system and to find effectiveness of implemented improvements.
- Chapter 7 discusses methods to improve translation quality. It explains ways to improve tagging accuracy of Urdu text which can improve factored translation. This chapter also introduces the idea of reordering of sentences before training.

- Chapter 8 explains experimental results. Effect of adding different documents into the corpus, effect of Large Language modeling and different variations of factored translations are explained in this chapter.
- Chapter 9 motivates for future research directions on the basis of implications of experimental results.

Chapter 2

Statistical Machine Translation

This chapter briefly describes history of statistical translation systems and then explains currently used methods of SMT. Section 2.1 explains the basic mathematics of statistical translation systems, section 2.2 explains word-based system and section 2.3 describes phrase-based systems which forms the framework for this thesis.

2.1 Statistical Modeling of translation System:

Modern-day machine translation systems originated in 1949, when Warren Weaver suggested applying statistical and cryptographic techniques from the nascent field of communication theory to the problem of text translation from one natural language to another natural language. But, initial efforts in this direction could not proceed because of doubts faced by this approach. There were doubts about building an automatic system to generate high quality translation. In 1966, the ALPAC (Automatic Language Processing advisory Committee) report stated that machine translation could not be preferred over low cost and low demand human translators. After this report, research in the field was almost stopped for a long period.

Rule-based and Knowledge-based system became the focus of research in this field. But, it requires lot of effort to design such systems, as they are based on linguistic patterns, rules and exceptions which are manually designed and hard-coded into the system.

Next advancement came in the early 1990's when bilingual corpora were made available in electronic form and statistical methods had been used in speech processing and multiple types of language processing. IBM Researchers proposed a model of statistical machine translation with the help of a probabilistic dictionary (Brown et el. 1990). Soon afterwards, automatic sentence alignment models were introduced by Gale and Church (1991) and Brown et al. (1991).

In 1993, Brown described a series of five statistical model of machine translation which are known as IBM Models. These models laid the foundation of word-based SMT. Models described the generation of translation of a sentence, word by word, using Shanon's (1948) noisy channel model of communication. Parameters of this models are estimated from the parallel corpus. Figure 2.1 shows how this model can be used for automatic machine translation.

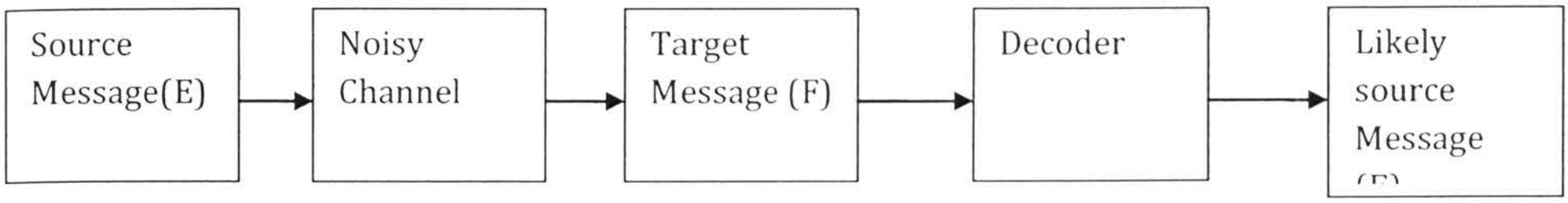


Figure 2.1: Translating with Shanon's noisy channel-model of communication

In this model, considering a French sentence F as an encoded English sentence implies that the probability of E being the intended translation of F can be expressed using Bayes's rule:

$$P(E|F) = \frac{P(E).P(F|E)}{P(F)} \qquad (2.1)$$

As the denominator does not depend on E, translation can be computed as maximum of numerator. Resulting equation of can be written as:

$$E' = P(E).P(F|E) \qquad (2.2)$$

The term P(E) represents the language model probability, and P(F|E) is the translation model probability. Language model probability is responsible for generating fluent English sentence and it is higher for a well-formed English sentence and it is independent of French sentence. Translation probability is the probability of sentence E as a translation of F and it is regardless of grammatical soundness which ensures that words are in their right positions. Equation 2.2 therefore assigns high probability to well-formed English sentence which also has high probability for translation from French sentence.

Translation Modeling Overview:

Machine translation is a difficult computational problem where we have to compromise for either time or complexity of the possible ways of computation. There are several strategies to tackle translation problem, but most of them are still out of reach.

Warren Weaver described languages as:

> "'[...] tall closed towers, all erected over a common foundation. Thus, it may be true that the way to translate from Chinese to Arabic [...] is not to attempt the direct route [...]. Perhaps the way is to descend, from each language, down to the common base of human communication –the real but as yet undiscovered universal language [...]" (Weaver 1949/1955)

Weaver viewed representation of an utterance into Interlingua as the fundamental step for ideal translation procedure. However, with the current state of natural language processing techniques, it is still far from reality. Other promising methods are based on semantic

transfer, where meaning of a sentence is inferred after the analysis of source sentence and then translation is generated from the inferred meaning of sentence. Due to lack of semantically annotated corpora, this technique cannot be implemented.

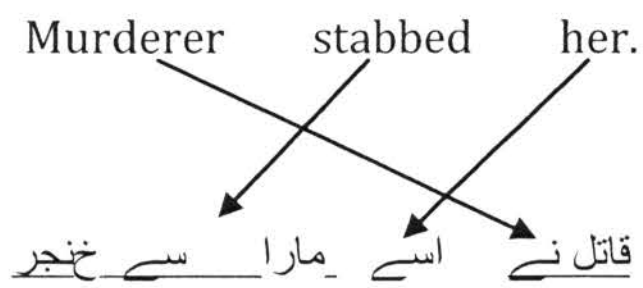
Next, there are syntax based approaches, as described by Yamada and Knight (2001), where source sentence is parsed to generate a syntactic tree. This tree is then transformed into target language syntactic tree and then surface form is generated from target language syntactic tree. Main obstacle in these approaches is the availability of parsing tools. Parsers and taggers are available for some widely spoken languages but still there are several languages which lack good linguistic tools.

Word-level translation is the simplest statistical approach to translation where translation is done word by word and afterwards, words are reordered to make sense in target language. It is the basic idea behind IBM models. The greatest advantage of word-level translation model is that it can be trained easily from the parallel corpus and does not depend on any linguistic information. However, drawback of such system is that, it fails to encapsulate syntactic and semantic information of source sentence and thus degrading translation quality.

2.2 Word-based Translation Modeling:

Mostly researchers are concerned about the posterior probability in the fundamental equation 2.2. Word based translation is also a way to model P(F|E).

A sentence cannot act as a unit to estimate P(F|E), due to sparseness. Therefore, sentences are break down into a sequence of words. In the IBM models, words in the target sentence are aligned with the words in the source sentence that generated them. The translation in Figure 2.2 contains an example of one-to-one and many-to-one word alignment.



Definitions:

Suppose we have E, an English sentence with I words, and F a French sentence with J words:

$$E: e_1, e_2, \ldots e_I$$

$$F: f_1, f_2, \ldots f_J$$

According to the noisy channel model, a word alignment a_j : j -> i aligns a French word f_j with English word e_i, that generated it (Brown et el 1993). Let A denotes set of alignments that covers all words in F.

$$A = a_1, a_2, \ldots, a_J$$

Therefore, from the product rule of probability, probability of a particular alignment A is the product of individual word alignment probabilities.

$$P(F, A \mid E) = \prod_{J=1}^{J} P(f_j, a_j \mid E)$$

P(F,A|E) is known as alignment model. Since there are several possible alignments over the set of E words and F words therefore to estimate probability of a particular translation, probabilities of all possible alignments are summed up.

$$P(F|E) = \sum_{A \epsilon\ A} P(F, A|E)$$

Translation Model Probability Estimation:

If the word order between source language and target language were same, word-alignment could have computed easily. But it is not the case, and word-alignments are learned from the sentence aligned corpora. Brown et el. describes a way to calculate alignments using Expectation Maximization (EM) algorithm (Dempster et el 1977). EM computes estimation of hidden parameters by maximizing probabilities over training data. Hence, its estimation of word-translation probabilities is done through selection of those word-alignments which maximizes sentence alignment probability in the training corpus. The closer training data is to the real-time data, the accurate will be the parameter estimation and eventually, the better would be translation. However, EM algorithm does not promise to give global optimum estimation of parameters.

For parallel corpora comprising of S aligned sentences, EM algorithm tries to find the optimum estimation of parameter θ, translation model parameter.

$$\theta = argmax \prod_{S=1}^{S} \sum_{A \epsilon \boldsymbol{A}} P(F, A|E)$$

Given the word alignment probability estimate θ, the IBM translation models then compute translation probability P(F|E) for a word-based SMT.

Drawbacks of word-based statistical translation models indicates the necessity of more sophisticated models with more parameters (e.g. fertility, probability of a word translating into more than one word) so that real-world translations can be handled. Drawbacks of

word-based SMTs are inefficient handling of reordering, null words and non-compositional phrases. Some of these issues were resolved by phrase-based translation systems which uses phrases as their basis.

2.3 Phrase-Based Translation Modeling:

In statistical machine translation, "phrase" does not mean a syntactic phrase; it stands for a segment of sentence comprising of more than one word. When alignments are performed on the phrases instead of words, local context information gets integrated into the translation model. Figure 2.3 shows an example of phrase-aligned English to Urdu translation, including one-to-one and many-to-one alignments.

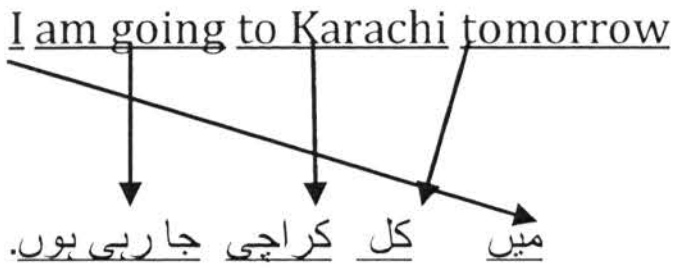


Figure 2.3: Example of Phrase-aligned translation from English to Urdu

Phrase level alignments have been an active research area in machine translation systems. One approach has been to use phrases corresponding to subtrees of a parsed sentence (Yamada and Knight 2001).

Phrase based translation models (Och 2003) use the automatically generated word level alignments from the IBM models to extract phrase-pair alignments. These extracted pairs are then used as the basic unit of translation. Kohen et el (2003) showed that phrase based translation system outperforms syntax based translation model of Yamada and Knight (2001).

A Simple Phrase based Model:

As described in the word-based model, phrase-based translation can also be model by noisy channel approach. Consider E is an English sentence and F is a French sentence which is segmented into I phrases.

$$F = f_1, f_2, \ldots, f_I$$

Since no linguistic information is used, all possible segmentations have equal probability. Each phrase f_i can be translated in to an English phrase e_i, using a probability distribution $\Phi(f_i|e_i)$.

Therefore, translation probability is

$$P(F|E) = P(e_1^l|f_1^l))$$

$$P(F|E) = \frac{P(e_1^l).P(e_1^l|f_1^l)}{P(f_1^l)}$$

The most probable translation sentence $\hat{E}$ is:

$$\hat{E} = argmax_E \prod_{i=1}^{N} \alpha_i^{x_{i(E|F)}}$$

$$= argmax_E P(e_l^l).P(e_1^l|f_1^l)$$

$$\hat{E} = argmax_E \prod_{i=1}^{N} P(e_i).\varphi(f_i|e_i)$$

Although, phrase based translation system extended word-bases translation system, using the same noisy channel approach. Phrase-based system tends to use log-linear model to compute translation probabilities.

Log-linear Model

It models P(E|F) as the combination of features x_i(E|F), these features are characteristic features of translation from F to E. If there are N features, then log linear model is represented as

$$P(E|F) = \frac{1}{Z} \prod_{i=1}^{N} \alpha_i^{x_i(E|F)} \qquad 2.3$$

Where Z is the normalizing constant and α_n is the weight assigned to the feature x_n(E|F). For a French sentence F, statistical machine translation system tries to find E which maximizes P(E|F), which can be written as,

$$\hat{E} = argmax_e \frac{1}{Z} \prod_{i=1}^{N} \alpha_i^{x_i(E|F)} \qquad 2.4$$

$$\hat{E} = argmax_e \prod_{i=1}^{N} \alpha_i^{x_i(E|F)} \qquad 2.5$$

If only two weights and features are considered and represented as language model P(E) and translation model P(F|E)

$$x_1(E|F) = log_{\alpha_1} P(E)$$

$$x_2(E|F) = log_{\alpha_2} P(F|E)$$

Then log-linear represents the basic equation of machine translation, and hence it is the special type of log-linear model.

$$\widehat{E} = argmax_e \frac{1}{Z} \prod_{i=1}^{N} \alpha_i^{x_i(E|F)} \qquad 2.6$$

$$\hat{E} = argmax_e \frac{1}{Z} \prod_{i=1}^{N} \alpha_i^{x_1(E|F)} \alpha_i^{x_2(E|F)}$$

$$= argmax_e P(E).P(F|E)$$

Log-linear model is more adaptable than traditional noisy channel model, as we can vary the weights according to the requirements of language pair. In practice, log-linear model is likely to express the language model and translation model probabilities as a composition of several features. The value of P(E) and P(F|E) could be expressed as combination of different features.

Log-Linear, Phrase-Based Translation Models:

As explained in equation 2.6, the most likely translation E in the general log linear framework is

$$\hat{E} = argmax_e \frac{1}{Z} \prod_{i=1}^{N} \alpha_i^{x_i(E|F)}$$

An advantage of log-linear model can be seen when taking log on both sides of equation 2.3 and then simplifying the notation

$$\hat{E} = -\log Z + \sum_{i=1}^{N} x_i(E|F).\log \alpha_i$$

$$\approx \sum_{i=1}^{N} x_i(E|F).\lambda_i$$

Probability of a translation is now the sum of the feature probabilities, so a single feature of zero will not skew the translation probability. Scaling the various features is done by setting the weights of features of λ_1 such that $\Sigma\ \lambda_1 = 1$ which allows EM algorithm to compute the probabilities of phrase-based models. Equation 2.7 can be written as,

$$\hat{E} = argmax_e \prod_{i=1}^{N} \alpha_i^{x_i(E|F)}$$

$$\hat{E} = \sum_{i=1}^{N} x_i(E|F).\lambda_i$$

Phrase-based log linear model by Kohen et el.(2003) includes several features to select the most likely translation. For instance, French phrases f_i are sequential within F but the English phrase $e_1,e_2,...,e_n$ might need reordering to make a grammatically correct sentence. This reordering probability is knows as distortion. It measures the distance between

English phrase and French phrase, which is marked as translation equivalents. Another important feature is δ to calibrate length of output sentence, as the model would usually prefer shorter sentence.

These parameters are estimated by Minimum Error Rate Training (MERT), a procedure which implements EM training step (Och. 2003). MERT operates by using a pre-calculated language model and set of probabilistic alignments, and then optimizing the weights for the feature to maximize the overall system's performance.

Chapter 3

Factored Translation Model

This chapter discusses factored translation and statistical model which is used to build factored translation model. Section 3.1 provided the overview of factored translation model and introduced the notion of factors. Translation in factored translation is explained in Section 3.2 and statistical model behind it is explained in Section 3.3

3.1 Introduction:

Phrase-based models calculate the translation as mapping small sequence of text. There is no linguistic information about the text and mapping is done on estimation of probability of surface forms. However, such information, morphological, syntactical or semantic, can play an important role in improving the translation.

Integrating such information could be beneficial because:

- If instead of surface forms, a more general representation of word is used to estimate probabilities then data sparseness issue can be overcome.
- Translation process can be better explained by morphological, syntactic or semantic level. If such information is available during probability estimation, it would certainly help in building better translation model.

Therefore, phrase-model was extended to incorporate additional information sources into the translation process. In this framework, a word can be annotated for multiple features and can be represented as a vector instead of as a token.

Factors:

A factor can be any other information which can be added to the surface form and can be helpful in the translation process. For instance, if the translation is being done from English to some morphologically rich language where words are inflected for number, gender and person, then translation system will not be able to formulate the required inflected form in the given context.

The word "blue" in Blue book will be translated as "نیلی" as the gender of book in Urdu is feminine plural, while it will be translated as "نیلا" for "Blue pen" as the gender of Pen is masculine.

Therefore it will be helpful to model between morphologically richer languages on the level of lemmas and then decide over the surface forms which are derived from the same lemma.

In this way, lemma and morphological information is translated separately and combine on the output side to generate the correct surface form.

3.2 Translation in Factored Model:

Factored translation model perceives translation as the generation of factored representation of output words from the factored representation of input words. Translation process is comprised of mapping of input factors to the output factors and generation of output factors from the existing output factors.

Factored translation follows the approach of phrase-based translation and work on chunk of text (phrases).But there are additional steps to translate vector of input factors into vector of output factors and to generate output factors from the existing output factors. Generation step works on word level and map output factors within individual words.

For instance, representation of read in English will be surface-form angry | lemma angry | part-of-speech VBD.

Mapping steps in morphological analysis and generation steps can provide the following output factors.

I. Translation: Mapping Lemmas
   angry -> ناراض | غصّہ
II. Translation: Mapping morphology
   VBD | sg -> VBD|sg|M , VBD|sg|F
III. Generation: Generating surface form
   ناراض |VBD| sg| F -> ناراض تھی
   ناراض |VBD|pl |M -> ناراض تھا

Application of these mapping steps to an input phrase is known as expansion. There are multiple options on each step; each input phrase can be expanded into many reasonable options. English word angry surface-form angry | lemma angry | part-of-speech VBD can be expanded as

1. Translation: Mapping lemmas
   { ?|غصّہ|?|?, ?| ناراض |?|? }
2. Translation: Mapping morphology
   {?|غصّہ|sg|F, ?| ناراض |sg|M , ?| ناراض |sg|M , ?| ناراض |sg|F }
3. Generation: generating surface forms
   { غصّہ تھی |غصّہ|sg|F , غصّہ تھا |غصّہ|sg|F, , ناراض تھی | ناراض |sg|M, ,
   ناراض تھا | ناراض |sg|M }

3.3 Statistical Model of Factored Translation:

Statistical modeling approach of factored translation follows the approach of phrase-based modeling. Major difference lies in the preparation of training corpus and type of learned model.

Training:

Parallel corpora for training have to be annotated with additional factors. If part-of-speech has to be used as a factor, then corpora have to be annotated with POS tags by using some automatic tagger. Next step is word-alignment and it is exactly same as phrase-based model. However, alignment can also be performed on other factors, such as lemma, instead of aligning on surface forms only.

Each mapping step forms a component of overall model. Translation and generation tables are learned from word-aligned parallel corpus and scoring functions, which helps in choosing among ambiguous mappings, are calculated.

In phrase-based translations, scoring functions are conditional phrase translation probability based on relative frequency estimation or lexical translation probability based on the words in the phrase. Similarly, models for translation steps are built from the word aligned corpus. For specified factors in the input and output, phrase mappings are extracted. These mappings are scored based on relative counts and word-based translation probabilities.

Generation steps are estimated for the target side only. They do not rely on word-alignments and additional monolingual data can be used to get better model. The generation model is learned on word-for-word basis. For instance, for generation of surface form from part-of-speech, a table with entries such as (مچھلی|NN) is built. Probability of this entry can be defined as, p(مچھلی| NN) and p(NN| مچھلی). These probabilities are obtained by maximum likelihood estimation.

Language Model for factored-based translation can be defined as n-gram language model over surface forms of words. Language model can also be defined over any sequence of factors.

Combination of Component

As factored translation models are viewed as the combination of several components (such as, Language model, reordering model, and translation steps). These components are defined as one or more feature functions which are combined in a log-linear model.

$$P(E|F) = \frac{1}{Z} \exp \sum_{i=1}^{n} \lambda_i \, h_i(E, F)$$

Z is a normalizing factor. Translation of input sentence f into the target language sentence e is break down into a set of phrase translations $\{(f_j,e_j)\}$. To compute the probability of a translation e given an input sentence f, all feature functions h_i will be evaluated.

For a translation step component, each feature function h_T is defined over the phrase pair (f_j,e_j), using a score function τ.

$$h_T\ (e,f) = \sum_j \tau(\bar{f}_j, \bar{e}_j)$$

For a generation step component, scoring functions Υ are defined over the target words e_k only:

$$h_G\ (e,f) = \sum_K\ \Upsilon(e_k)$$

These feature functions are obtained through scoring functions (τ, Υ) during the training of translation and generation tables. Feature weights λ_i of log-linear model are estimated through minimum error rate training method.

Decoding:

Instead of looking into phrase table only, multiple tables have to be looked up and their content has to be combined. Since all mapping steps operate on same phrase segmentation, expansion of these mapping steps can be efficiently pre-computed prior to the beam search and stored as translation options.

For a given input sentence, all possible translation options are thus computed before decoding.

Further enhancements:

Use of Automatic word classes:

Automatic word classes are also used as a factor during traning, these classes are obtained by clustering words based on their contextual similarity.

Integrated Recasing:

Training data is lower cased to generalize over different cased surface form which necessitates a post-processing step to restore case in the output. In factored model, it is possible to integrate this step into the model by adding a generation step.

Chapter 4

MT Error Evaluation

In this chapter, we explain the manual annotation scheme which is used to annotate MT output. We also discussed examples from the annotated test set. Section 4.1 presents the overview of related work which is done previously in this direction. A brief discussion of error classification and explanation of different error types are presented in Section 4.2, and in Section 4.3 examples are given from the annotated test set, Section 4.4 presents counts of different types of errors and summarizes the result of analysis.

4.1. Related Work

Evaluation of MT output plays an important role in the determination of the performance of the system. Edit-Distance based measures, word error rate, WER, and position independent error rate, PER, and n-gram based measures, BLEU (Papineni et al., 2002) and NIST(Doddington,2002),are among the most popular methods to evaluate statistical machine translation systems. In these techniques, quality of translation is determined by finding the similarity between human translated text and machine translation system's output. Human judgment is also used to evaluate translations. In both of these methods, a numeric value is assigned to the text which represents quality of translation. However, in order to improve translation quality, one has to analyze text and investigate types of errors, so that emphasis can be made on the most frequently occurring errors.

Since the inception of the machine translation, several methodologies have been defined to classify errors. Falagan [et al. 1994] classified errors into different types and ranked error types based on its effect on the intended meaning of the original text. Vilar [et al. 2006] used a hierarchical framework for error classification and it is one of the more detailed classifications till date. In this scheme, errors are first classified into five major categories and then subtypes are defined to encapsulate finer details of errors. Figure 4.1 shows the diagrammatic representation of this error hierarchy.

4.2. Annotation Scheme:

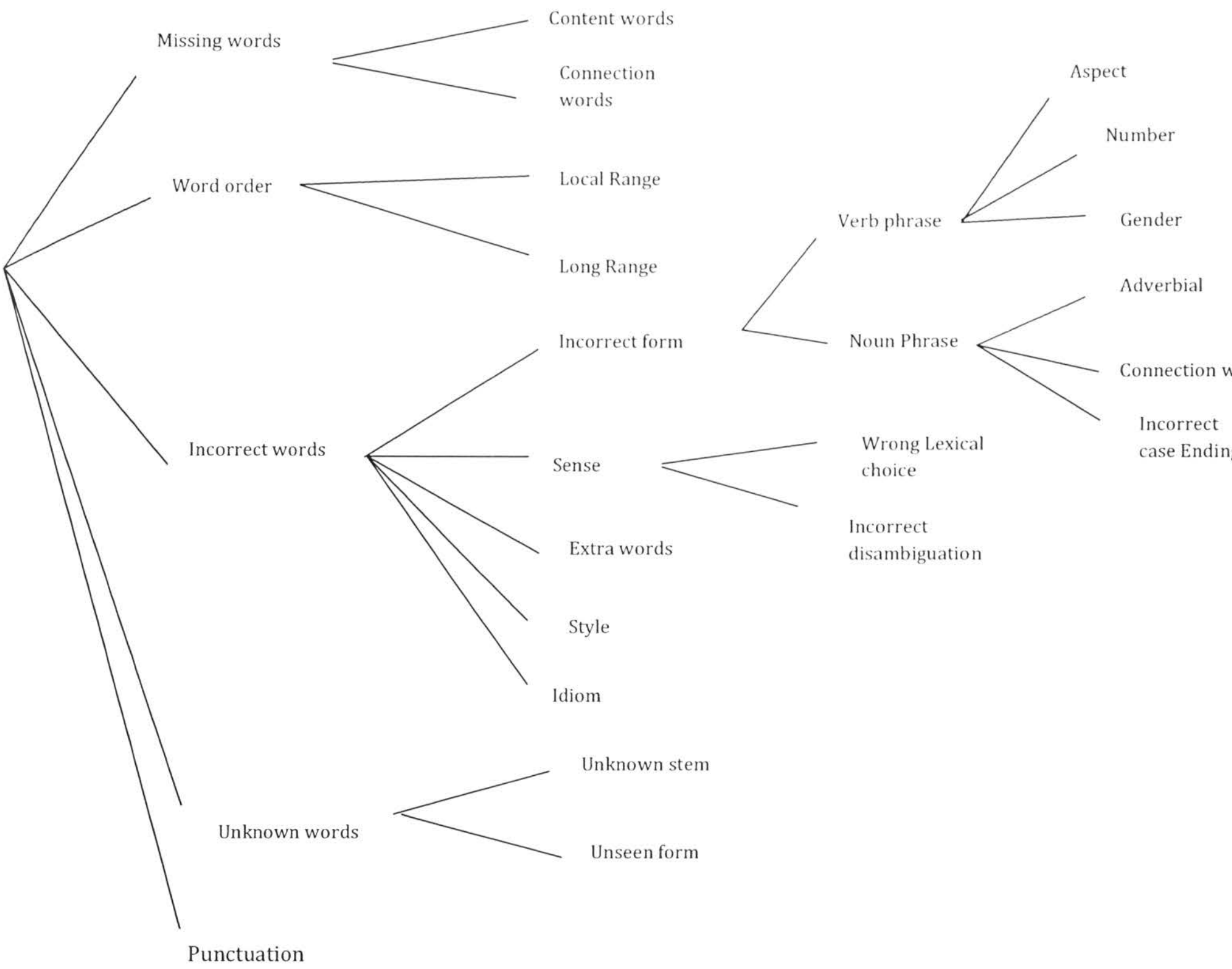
For the analysis of the current system, Vilar's hierarchy of errors is used with certain changes in order to specify certain issues related to free word order languages.

Explanation:

Following is the explanation of the scheme.

Missing Words:

"Missing words" indicates the absence of a word in the generated text. It is divided into two classes, missing content word and missing connection word. Content words are those words which are important for determining meaning of a sentence while connection words are those words which define the semantic relation between content words.



Word Order:

This type of error occurs when system places words in places which are not coherent with target language word order. Urdu is a relatively free word order language, where there must be an order insider a phrase while phrase can occur at any place in the sentence.

Word order is further classified as long ranger word order and local word order.

Local range word order indicates the improper word order within a phrase. e.g.

King of Mankind بادشاہ کا لوگوں

In the above noun phrase," کا" should be place after "لوگوں"

Long range word order is the error where either a content word or a connection word is not placed in its proper phrase.

Incorrect word:

This category of errors encompasses a wide area of different types of errors; first sub category, incorrect forms, indicates presence of errors which are caused when words are not inflected in order to have agreement with other phrases.

Verbal phrase is modified according to the number, gender and person of the subject or object of the sentence.

I read a book.	میں نے کتاب پڑھی
I read a novel	میں نے ناول پڑھا

Here, verb "read" is inflected because of gender of object (In Urdu, book has feminine gender while novel has masculine gender).

Noun phrase is inflected according the tense of verb.

I go to the market	میں بازار جاتا ہوں
I was made to go the market.	مجھے بازار لے جایا گیا

In the above two sentence, form of the word "I", is changed (میں and مجھے) because of aspect of verbal phrase.

There is also an agreement of number, gender and number with the head of noun phrases.

lrig doog	اچھی لڑکی

Adjective "good" is inflected for the feminine gender of girl.

Players of Pakistani Cricket team	پاکستان کی کرکٹ ٹیم کے کھلاڑی

Here, connection words," کا کے کی" are used according to the gender of the next word.

Second subcategory deals with errors caused by different incorrect forms of words. In verbs, it is related to errors caused by usage of incorrect tense, aspect or person while in other phrases, it can occur due to improper form with respect to main word.

Extra words:

This category indicates the presence of errors caused by the presence of extra words generated by the MT system.

Sense

This subcategory indicates errors presence of errors which occurred when system is not able to identify the correct sense of words or when it cannot disambiguate a word.

Punctuation

Punctuation errors, caused due to misplacement or absence is marked by this category.

4.3. Examples

Following are some examples from the annotated output of baseline MT engine and their explanation.

SRC: (Ever since He has chosen you,) your Lord has not forsaken you. Nor is He displeased (ever since He has taken you as His Beloved).

REF: بنایا محبوب کو آپ سے جب) ہی نہ اور چھوڑا نہیں کو آپ (ہے فرمایا منتخب کو آپ سے جب) نے رب کے آپ ہے ہوا ناراض (ہے

TST: (تب sense: wrong lexical choice ہے لیا فرما منتخب تمہیں نے اس سے extra word!: punctuation) نہیں رکھیئے چھوڑے تمہارے نے رب تمہارے -:punctuation مکروہ نہ اور sense::Wrong lex choice جانا (تمہارے وہ بھی تب :missing connection word کیا اختیار verbphrase:aspect ہے جس ہے (حبیب اپنے طرح

In this translation,"has not forsaken" is translated as "نہیں رکھیئے چھوڑے" which is wrong with respect to person and tense of the sentence. Negation mark is placed in the end of the sentence instead of after main verb.

Second error in the test sentence is the wrong sense of word "displeased" which is occured in the infinitive form in the sentence instead of past indefinite. "as his beloved" should be translated as "طرح کی حبیب اپنے", however "as" can be translated into different ways depending on the context, system found out the right sense but failed to find the correct connection word over here which destroyed the intended meaning.

Further more, "Ever " should be translated into "jab " and it indicates a point of time in future while "tab" ,can be translated as "since", indicates a point of time in the past. MT system could not disambiguate between the two and used the wrong word.

SRC: By the fig and by the olive!

REF: قسم کی زیتون اور قسم کی انجیرo

TST: قسم کی قسم کی انجیر word order:local range اور زیتون!:punctuation

For smaller sentences, MT system worked fine except some word order issues. Here the phrase, "by the olives " got translation word for word and was not reordered according to the Urdu grammatical structure.

SRC:Whether he (the whispering Satan) comes from the jinn or mankind.
REF: وہ خواہ (شیطان انداز وسوسہ) جنات میں سے ہو یا انسانوں میں سے
TST:وہ خواہ (شیطان پھوسی کانا) لوگو یا word form:incorrect case ending جنات سے آنے گا verb phrase: aspect -

In this examples, literal translation of "comes" is present, which is acceptable for the machine generated translation while for the reason of stylistic translation, meaning of "comes" is not present in the sentence. This poses a question that whether it should be classified as error or it should be assumed as acceptable translation of the source sentence.

When noun or participle acts as adjectives for the main noun, an extra word is used in the Urde sentence to define the relation between two nouns or participle and the main noun. Generated output is missing this extra word which is here,"والا کرنے".

SRC: . Surely We sent down this (Holy Quran) during the Night of Destiny

REF: بے اتارا میں قدر شبِ کو (قرآن) اس نے ہم بیشک

TST: . ہم بیشک missing:connection word گئے اتارے : incorrect form: aspect اور :extra word

اس (قرآن) :missing connection word رات - :punctuation جب : تقدیر کے:incorrect connection word دوران sense: wrong lexical choice

4.4. Analysis of Annotation

Analysis of the translation reveals that there two main sources of errors , one is the incorrect formulation of phrases, because of incorrect case markers form and incorrect order within a phrase, in verbal phrase it is because of incorrect tense ,aspect and modality. Second main source is the missing connection words.

Both of these issues could be because of the size of corpus and learning is compromised because of limited availability of parallel data.

Type	Count
Missing Words	45
Content word	28
Connection word	17
Word Order	23
Local Range	18
Long Range	5
Incorrect words	74
Incorrect form	59
Verb phrase	22
Aspect	22
Number	18
Gender	10
Noun phrase	37
Adjectival	15
Connection word	12
Incorrect case ending	10
Extra words	0
Sense	15
Wrong Lexical choice	10
Incorrect disambiguation	5
Style	-
Idiom	-
Unknown words	-
Unknown Stem	-
Unseen form	-
Punctuation	5

Table 1 Error Counts

Chapter 5

Data Acquisition

Parallel corpus is the main asset in the development of statistical machine translation systems. Quality of translation depends on the domain of parallel corpus, degree of parallelism between the pair of language, style of writing which is used in the text. System trained on news corpus will not give good translation of academic text.

Essence of phrase based translation is in the determining proper word alignment between source language and target language sentence. Accurate word alignment can only be determined when corpus is properly sentence aligned and when difference in the word order of two languages is not much. Available tools for aligning sentences only work when sentence i in source language is mapped to some sentence i+1 sentence in target language and it fails when text is not placed in the same order in two languages.

5.1 English-Urdu Parallel Data Acquisition:

Unfortunately, there is not a lot of parallel electronic data available on the web, only two available parallel corpus are: EMILLE and CRULP's data

EMILLE:

EMILLE project, by the University of Lancaster, is one of the initial efforts to make Urdu corpus available for research. Project has 200,000 words of English text translated in to Urdu but the quality of translation is not good. Moreover, this corpus comprised of text from interviews of different personalities, therefore it is a corpus of spoken Urdu which, at times, mixed up with lots of English words.

CRULP:

Recently CRULP released Urdu translation of 6214 sentences of Penn tree bank, having 153,611 words. Corpus is related to news domain and is the only reliable resource available.

CRULP also released POS tagged data of 4k sentences.

Quran:

Islamic holy book Quran, originally in Arabic, is translated into different languages. According to Wikipedia, it is available in 102 languages. It contains 6236 verses. Different translators translate it into English and Urdu. Among the famous English translators are Yusuf Ali, Marmaduke Pickthal, Dr. Mohsin, Dr. Abdul daryabadi. In Urdu, there are translations from Ahmed Ali Lahori, Jalandhri, Usmani.

However, different translators have different style of translating the same Arabic text; therefore other translations are combined to get an extensive parallel corpus. Three English translations and four Urdu translations were added to the corpus.

Sahih Bukhari and Sahih Muslim:

Another religious text, which is known to be available in many languages, is Hadith books. Sahih Bukhari and Sahih Muslim are available on web in English in several websites; Urdu version was obtained from web in Inpage format and was converted into Unicode format.

But there were differences in the presentation of same text in two languages; Urdu version has chapters which were encompassing broader topics while all English version has fine grained chapter titles. Moreover, there were some extra texts (explanations) in English version, making it difficult for hunalign.

Urdu text was manually divided into smaller units and then hunaligned. But alignments obtained still were not good. Inclusion of this data into the corpus decreases the BLEU score.

Urduseek.com

There is good English to Urdu dictionary available at urduseek.com. All the html pages of dictionary were downloaded and html tags were removed, to include it into the corpus.

Monolingual Data from Blogs for Language Modeling:

Urdu syntax is completely different from English. For a better modeling of target side syntax, Urdu blog's data was downloaded to build a better language model.

Chapter 6

Experimental Framework

In this chapter, we describe the experimental setup used to build and evaluate phrase based translation system. Section 6.1 explains the choice of data set as well as the steps which are used to process the data. The tools used for creating a translation system are mentioned in section 6.2 and Section 6.3 describes the processes for tuning and evaluation of system.

6.1 The Corpora

Due to unavailability of English-Urdu parallel corpus, we searched web to get parallel texts. Famous religious books, which are available in both languages, are used to create parallel corpus. Details of data acquisition are discussed in chapter 7, where we explained all the data sources which were used as well as their quality. Corpus consisted of 25K lines of text and a dictionary of 226348 words. Dictionary is created by crawling on-line English to Urdu dictionary website which has 128997 English-Urdu word pairs and 97381 Islamic names written in both language. Development set comprised of 800 sentences, while test set consisted of 200 sentences.

Preprocessing of Corpus:

There were different versions of same punctuation symbols used by different documents in Urdu. All such symbols were standardized to have same symbol.

There are two different versions of sentence marker '۔'[Unicode value 1748] and '-' [Unicode value 45], it was standardized by transforming all into '-', preceded by a space. Space was inserted because when sentence marker is placed right after a word (e.g. گا۔), it is treated as a part of word and not as an punctuation symbol.

Arabic symbol," ö" is used in traditional Arabic literature. It was present in the Urdu translation of Quran to mark the end of a verse while English translation does not have this marker. It was removed from the text.

Different forms on inverted commas " and ‘‘ are converted to one single notation.

Multiple numeric representations, " 1."," (۱)" and ".۱۹" are converted to one notation and English digits are replaced by Urdu digits.

Quite often, an elaborative translation was found to be within brackets in Urdu text. This is done to follow traditional style of writing. But the corresponding English text was not found in brackets in the same place. Therefore, brackets were removed from the Urdu text.

Factored Corpus:

We selected part-of-speech and four-lettered stem as factors. Stanford Maximum Entropy tagger (et el Toutanova and Manning. 2000) was used to tag the data. For English,left3word-wsj model accompanied with tagger, is used to tag the corpus while for Urdu model was trained on CRULP's available tagged corpus of 4k lines.

Stanford parser also performs sentence segmentation by looking at broader context. If it will be used from the command line then the number of tagged lines in English would have been different than the number of lines in Urdu. Therefore, a wrapper Java class was written which tags sentences one by one and does not take broader context in account. A simple perl script was used to create four-lettered stem of each word.

Using the tagger and stemmer, we produced corpus formatted for training, wherein each word is expanded to a feature bundle surface-form|POStag|stem, as we can see below:

the|dt|the beneficent,|nnp|bene the|dt|the merciful.|nnp|merc
master|nnp|mast of|in|of the|dt|the day|nn|day of|in|of judgment,|nnp|judg

۲|cd|۲ بڑا|jj|بڑا مہربان|jj|مہرب نہایت|i|رحم والا||رحم نہای|wala| والا
۳|cd|۳ انصاف|nn|کے انصا|cm|دن کے|nn|کا دن|cm|حاکم||حاکم کا

6.2 Software tools:

We used standard phrase based statistical machine translation framework for our experiments, along with the following tools.

Moses : Translation models are created by Moses. Moses uses a multi-word phrase translation table along with the language model to translate sentences.

GIZA++ : Word alignments are done by GIZA++, which is an implementation of IBM Models (Och and Ney 2000)

SRILM Toolkit: The n-gram and factored language models were trained using SRI Language Modeling toolkit (Stolcke 2002)

MERT: The translation system tuning was done using the Minimum-Error-Rate Training tool, which is an implementation of the Expectation-Maximization(EM) algorithm described in Section 2.3. MERT operates by using a pre-calculated language model and set

of probabilistic alignments, and then optimizing the weights for the features to maximize the overall system's BLEU score on a reference set.

6.3 Tuning and Evaluation

Translation Evaluation

There are two ways to evaluate machine translation system; Manual Evaluation, which scores sentences based on fluency of sentence and accuracy of content and Automatic Evaluation, where translation is judged by calculating similarity with reference translation. While generally agreed to be the most desirable method for translation evaluation, manual translation is too time-consuming and expensive to be used to compare many different versions of a system during development and testing.

The standard automatic evaluation is BLEU, proposed by Papineni et al. (2002). BLEU is a precision-based metric that compares the system output and reference translations using n-gram matches between the sentences. While Callison-Burch et al. (2006) show that BLEU at times can have a weak correlation with human judgments on machine translation output from heterogeneous system types, it is still sufficient for tracking the results of similar systems, as is done here.

Tuning

In log-linear statistical translation system, most probable translation is the one which maximizes the product of several weighted feature scores, as described in Section 2.4. Parameters of this model significantly affect translation quality, as they guide the decoder through the space of possible translations. These parameters are learned using Minimum Error Rate Training (MERT) (Och. 2003). MERT iterates over a test set and considers rescoring translations with different parameter weights until the Bleu score of the test set has converged. We tuned the translation systems on a set of 800 sentences.

We judged the experiments on the provided evaluation dataset which is disjoint from both the training and development sets in order to prevent over-fitting the system to the supplied data. As such, we used the system scores on the development set as rough estimates, and only ran the experiments on the evaluation dataset after all system tuning was complete.

Chapter 7

Searching for better Phrase based Machine Translation

In this chapter, we explained different techniques which can be helpful in improving the translation of phrase-based machine translation system. Proposed techniques are based on error analysis which is described in Chapter 4. Section 7.1 points out the main sources of errors and gives motivation to address these issues to get a better MT engine. Related work in this dimension, specifically to improve translation from English to Urdu is discussed in Section 7.2. Section 7.3 discusses usage of case markers to improve tagging of noun phrases during the PoS tagging of Urdu corpus. Section 7.4 discusses reordering of sentences on the basis of dependency parsing.

7.1 Reasons of Errors:

There are several reasons of not getting very promising results for English to Urdu statistical machine translation. Some of the issues which we realized after analyzing baseline system are following.

Scarcity of good linguistic resources: There is not any large English-Urdu parallel corpus available and therefore Moses cannot train the MT system to encapsulate different linguistic aspects of English and Urdu. Due to lack of tagged Urdu data, automated taggers cannot efficiently learn to tag Urdu.

Secondly, Urdu is morphologically rich language as compared to English and therefore it is difficult to generate appropriate word forms and case markers on the target side, especially when there is a limited quantity of parallel corpus.

Third reason is that, Urdu is subject-object-verb language while English is subject-verb-object language and when phrase based systems are used between languages with very different word order, long distance reordering becomes one of the key weaknesses.

7.2 Ways to Improve translation:

There are multiple ways to improve output of statistical machine translation system. These techniques fall into two categories. One of them is handling morphological variations on either source side or target side and the other is handling syntax by integrating syntactic information in to the machine translation system.

Modeling target side syntactic structure:

Better modeling of syntactic structure can also improve translation. It can be done by POS factored translation model. Figure 7.1 shows proposed translation model.

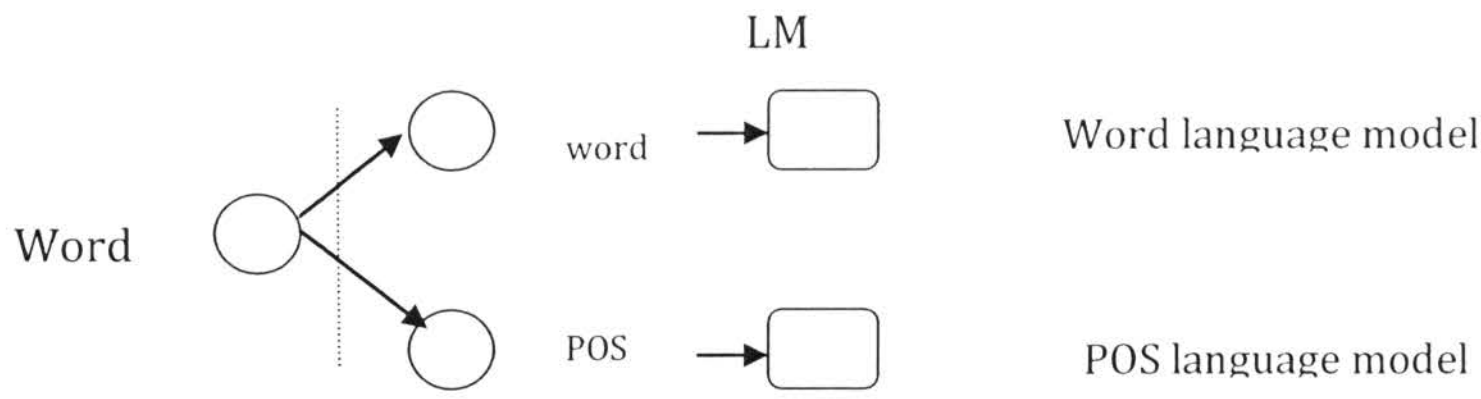


Figure 7.1 POS factored model, source word determined target side factors (target word, target word POS) each factor is associated with its language model.

Translation probability in this model is calculated as a log linear combination of phrase probabilities, reordering model probabilities and each of the language model probabilities.

$$P(e|f) \sim p_{lm-word}(e_{word}) * p_{lm-POS}(e_{POS}) * \sum_{i=1}^{n} p(e_{word-j}, e_{POS-j} | f_j) * \sum_{i=1}^{n} p(f_j | e_{word-j}, e_{POS-j})$$

$P_{lm\text{-}word}$ represents language model probability which is estimated on the surface form and $P_{lm\text{-}POS}$ denotes language model probability over part-of-speech tags. $p(e_{word-j}, e_{POS-j} | f_j)$ and $p(f_j | e_{word-j}, e_{POS-j})$ are translation probabilities.

Urdu Tagging:

In order to get better results for the described model, Urdu data should be properly tagged by using some statistical tagger. Unfortunately, there are only 4k lines of text available for training POS tagger. To get better results from tagging, a number of heuristics are applied to tag words correctly.

A hybrid approach is used to tag Urdu corpus. Initially Stanford POS tagger is used to tag Urdu corpus and afterwards it is processed to correctly tag nouns by identifying case markers. Another advantage of doing so, is to properly tag case markers so that MT training should learn the placement of case markers

Case Markers:

Urdu is a relatively free word order language which allows phrases to occur at any place (except verbal phrase which usually occurs at the end of sentence) in the sentence but there is a strict order of words with in phrase.

Verbal phrases consist of verb along with helping verbs, while noun phrases are constituted of noun and a case marker. Case is a system of marking dependent nouns for the type of relationship they bear with their head [Ahmed et.el., 07]. e.g. subject and object of a verb can be marked by a case.

There are eight cases in Urdu: nominative, accusative, dative, ablative, instrument, genitive, locative and vocative [Ahmed et.el.,07]. Since English is a fixed order language, the subject and object are distinguished by their positions.

Example of case markers:

میں نے ربیع کو ایک کتاب دی

اسکول کو مرمّت کی ضرورت ہے

Case	Urdu
Nominative	ϕ
Ergative	نے
Accusative	کو
Dative	کو
Instrument	سے
Ablative	سے
Locative	میں،پر

Since case markers are closed class words and can easily be identified in a sentence, we can tag its previous token as NN .In this way, we can also mark case markers accurately.

Reordering:

Quality of translation can be improved when word alignment is accurate and it is possible when placement of words in source and target language is similar. If the word order requirement of two languages has major differences then reordering decision is very difficult to take based on statistical information due to dramatic expansion of the search space with the increase in number of words involved in the search process.

Related Work:

Over the last few years, different reordering schemes have been proposed. Distance based reordering model (Koehn 2003) is implemented in Moses. It penalized non-monotonocity by applying a weight to the number of words between two source phrases corresponding to two consecutive target phrases. (Tillmann 2004; Koehn et el. 2005,Al-Onaizan and Papineni, 2006) extended this model with lexicalized reordering by applying different weights to different phrases. Hierarchical phrase reordering model is introduced by (Galley and Manning 2008) which determines phrase boundaries using shift-reduce parsing. However, none of these models change word alignments done during the training of SMT systems therefore word-alignment errors are not rectified.

Apart from adding different weights and incorporating different models into the basic SMT model, there is another approach to solve reordering problem. It is done by putting syntactic analysis of target side into both modeling and decoding. Constituency trees (Yamada and Knight,2001; Galley et.al.,2006;Zollman et. al., 2008) and dependency trees (Quirk, et.al., 2005) were shown to give significant imp9rovements in the translation quality. Hierarchical phrase-based approach (Chiang, 2005; Wu, 1997) showed good results for Chinese to English machine translation.

Researchers have also tried to do syntactic analysis of source side and then to use this information in the reordering of target side. Collins et.al., 2005 used manually designed rules to reorder German sentences and showed that this approach can increase BLEU score from 1 to 2% over baseline.

Fei Xia et.al., 2004 described a way to automatically extract reordering rules for French to English translation. Rules are extracted from the parsed tree and results show significant improvement on the BLEU score over the base line system.

Reordering Model in Urdu

Xu et.et., 2009 did experiments on translating from subject-verb-object SVO language to subject-object-verb language. Translation from English to five SOV languages, including Urdu, was considered. Reordering was done using manually designed rules and applied after dependency parsing of sentences. Applied rules were related to verb placement, adjectives, nouns and prepositions. Rules were extracted by analysis of Korean sentences and were generalized for all five languages. However, mentioned reordering of adjectival phrase is not applicable for Urdu.

We tried to find reordering rules for Urdu, especially for verbal phrase, prepositional phrase, question sentence and negative sentences using the same approach as mentioned above.

Dependency Parsing:

Dependency is based on the idea that syntactic structure of a sentence consists of asymmetrical relations between the words of the sentence. A dependency relation holds between head and dependents.

A dependency parser parses a sentence to identify grammatical relation between words. We used Stanford statistical dependency parser to find the dependency relations and then we apply reordering rules.

A quick brown fox

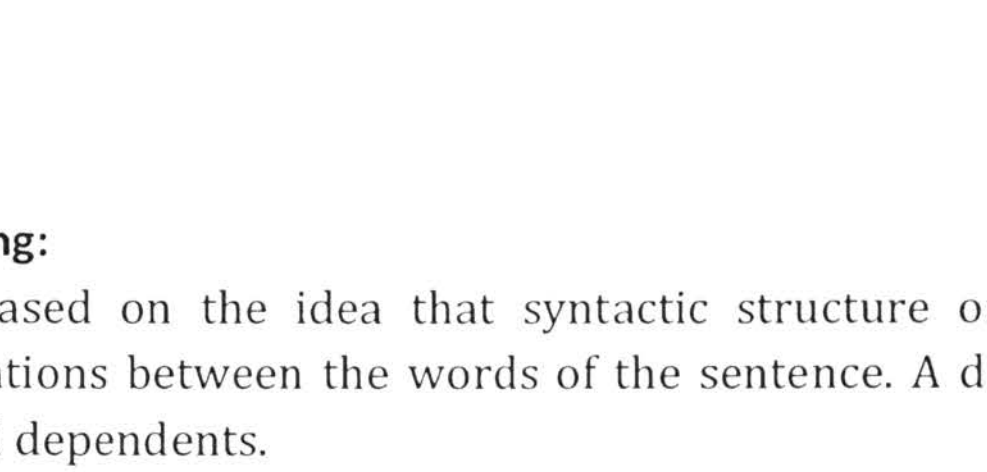
det(fox,a) amod(fox,quick) amod(fox,brown)

Verb Phrase:

In Urdu, main verb comes at the end of the sentence. In order to obtain this structure, dependency structure of verbal phrase is swapped, and if dependent is the head of any other structure its position is also changed accordingly.

English Sentence: She has been writing a novel since last month.

Dependency Structure:



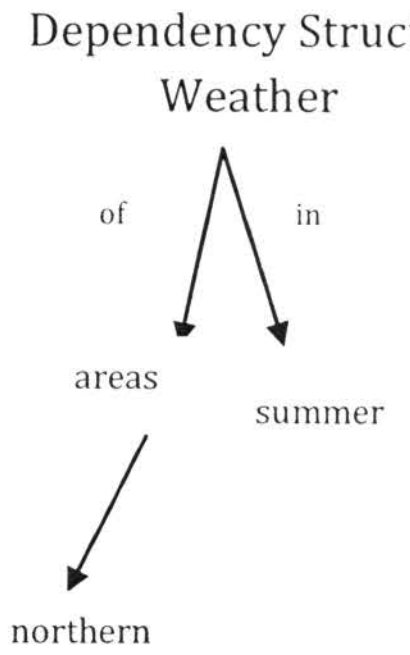
Reordered English Sentence: She last month a novel writing has been

Urdu Sentence:وہ پچھلے مہینے سے ایک ناول پڑھ رہی ہے

Prepositional Phrase

Preposition comes after nouns in Urdu, therefore dependency structure is also swapped here

Weather of Northern areas in summer



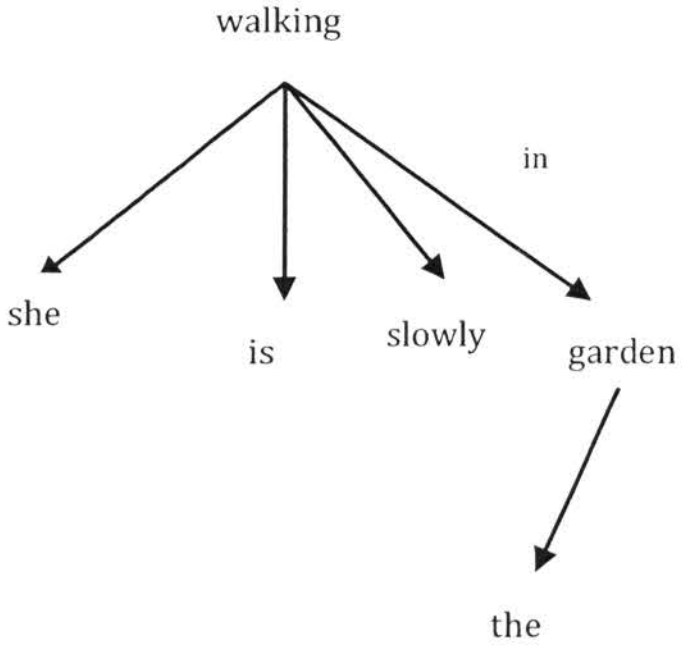
Reordered Sentence : Nothern areas of summer in weather

Urdu Sentence: گرمیوں میں شمالی علاقوں کا موسم

Adverbial Phrases:

English Sentence: She is walking slowly towards the garden.

Dependency Structure:



Reordered Sentence: she slowly towards the garden is going

Urdu Sentence: وہ آہستگی سے باغ کی جانب جا رہی ہے

Chapter 8

Results

This chapter presents results of experiments which are done to analyze the effectiveness of different implemented methods to improve the translation of phrase based machine translation. Section 8.1 presents the outcome of experiments with different corpus. Section 8.2 discusses the translations generated by factored translation model and effect of dependency based reordering of source language sentence on the result of translation is discussed in Section 8.3

8.1 Experiments With Different Corpus:

Due to unavailability of corpus, parallel documents from multiple sources are collected and combined to create a reasonable size parallel corpus for training of Moses. To find the effectiveness of adding a particular document into the corpus, we did experiments by incrementally adding these parallel documents.

With only Quranic data

Initially, we trained a small experimental system with Quranic translation. Since the parallel corpus consisted of only verses, there was no need for any preprocessing and sentence alignment.

Training data	5909
Dev data	300
Test data	100

Table. 8.1 Size of training, development and test data sets

NIST score	2.779
BLEU score	0.144

Table 8.2 shows the calculated BLEU and NIST score on the test data, trained on Quranic text.

As described in chapter 5, there are multiple translations of Quran in English and Urdu and different translators have different style for translating the same Arabic text. Since probability of sentence alignment in religious script is higher, we decided to add various English and Urdu translations in to the corpus.

Resultant corpus comprised of 20708 lines, development and test data was not changed. BLEU score on the test data increased from 0.144 to 0.179 which indicated that with the increase of well-aligned corpus quality of translation can improve.

Training data	20664
Dev data	300
Test data	100

Table. 8.3: Size of training, development and test data sets with multiple translations of Quranic Text

NIST score	4.1538
BLEU score	0.1793

Table 8.4 shows the calculated BLEU and NIST score on the test data, trained on Quranic text.

With Dictionary

We described English-Urdu dictionary in chapter 4, it contains 226378 word-meaning pairs. This dictionary is created by crawling a urduseek.com

Online dictionary has the following format

ab initioسرے سے ۔ آد سے ۔ شروع یا ابتدا سے ۔ اول سے

In this format, multiple Urdu words corresponding to English word are represented together, to make it processable by Moses, such single entries were converted in to multiple one.

ab initio اول سے

ab initio شروع یا ابتدا سے

ab initio آد سے

ab initio سرے سے

Training data	20664 lines and 226378 dictionary word
Dev data	300
Test data	100

Table 8.5 Size of training, development and test data sets with multiple translations of Quranic Text

With inclusion of the dictionary into the corpus, BLEU score on the test data set increased by 0.45.

BLEU Score	4.1821
NIST Score	0.1835

Table 8.6 shows the calculated BLEU and NIST score on the test data, trained on Quranic text and dictionary.

Reason of increment in the score can be due to the better translation model by the inclusion of dictionary.

With Tafseer

Parallel text of Tafseer which consisted of 4k lines is also a good resource, but there was not one-to-one correspondence between English and Urdu text. Inclusion of this book into the corpus decreased the BLEU score and it became 0.159

BLEU Score	0.159
NIST Score	4.15

Table 8.6 shows the calculated BLEU and NIST score on the test data, trained on Quranic text, tafseer and dictionary.

Effect of Large Language Model:

Language model which is built on large monolingual data can be helpful in learning target side syntax and can improve the translation. Therefore, we gathered Urdu data from multiple Urdu blogs and refined it by removing the HTML markups.

Training corpus	29312
Monolingual corpus	545389
Dev Data	800
Test Data	200

Table 8.7 Size of training, development and test data sets and Size of monolingual data for language model

NIST score	4.2774
BLEU score	0.1617

Table 8.8 shows the calculated BLEU and NIST score on the test data, trained on parallel corpus with large language model.

BLEU score confirmed that using a large monolingual data increased the quality of translation.

8.2 Experiments with Factored translation:

We conducted various sets of experiments in the factored translation model. Initially, we tried to find out the factors and statistical models which can give the best translation. In order to find it out, we trained statistical translation system with different settings of parameters and factors.

Secondly, we tried to find the effect of different ways of part-of-Speech tagging on the output of factored translation. We also did experiment of training factored translation system with translation factors comprising of surface form and PoS tags along with language model of surface form and language model of PoS tags.

For the initial experiments, which were conducted to investigate the effect of different models and factors on translation, we select a subset of 10k lines of parallel corpus. Development set consisted of 800 sentences and test set comprised of 200 sentences.

To find out the impact of various models on the translation system, we started factored translation system with simple models and then incrementally added other factors and models.

To evaluate translation quality, we first built translation model with translation factors consisting of surface forms. This provides a baseline against which to compare the performance of translation system using other models.

Model	Dev BLEU	Test BLEU
Baseline Factored translation model	0.0883	0.0782

Table 8.9 Development and test BLEU score for Baseline factored translation system

Table 8.10 shows the BLEU score of translation system using lexicalized reordering model on surface form. The scores are nearly identical to the baseline in table 1, differing by 0.002 on the development set and 0.001 on test set. This shows that reordering model does not have any significant impact on the quality of translation.

Model	Dev BLEU	Test BLEU
With Lexicalized reordering model on surface form	0.0881	0.0781

Table 8.10 Development and test BLEU score for factored translation system with reordering model on surface form.

Next we performed experiment with alignment model and defined stems of word as the alignment factor and BLEU score increased by 0.571.

Model	Dev BLEU	Test BLEU
With Lexicalized reordering on surface form and Alignment model based on stem of word	0.145	0.143

Table 8.11 Development and test BLEU score for factored translation system with reordering model on surface form and alignment model on stems.

Impact of Generalized tagging for Urdu:

Urdu data was tagged with Stanford tagger which was trained on the tagged corpus of 4k lines [66k words]. Since the training data was not enough, the accuracy of tagger on test data which consisted of 300 words was 57.7%.

To improve the accuracy, we decided to map Urdu PoS tag set on a smaller set of tags and to convert fine grained PoS tags into their general representation. It was done to get translation system learn the SOV structure of Urdu. By doing so, accuracy of the tagger increased to 70% and by marking nouns, as defined in section 7.2, appearing after case markers, accuracy of tagger goes to 73%.

To compare the impact of improved tagging, we built translation models with and without generalized tagging. Translation factor was defined to map input surface form to output surface form and tag.

Model	Dev BLEU	Test BLEU
Without generalized tag	0.03	0.028

Table 8.12 BLEU score of Translation model with Urdu Pos factor without reduction

Model	Dev BLEU	Test BLEU
With generalized tag	0.04	0.031

Table 8.13 BLEU score of Translation model with Urdu Pos factor with reduction

Experiment with Reordering:

Experiment with reordering was done with same data set comprising of 20k sentences. Sentences were parsed with Stanford parser and reordered according to the rules defined in section 7.2. There is a increase one BLEU point compared to the translation system without reordering.

Model	Dev BLEU	Test BLEU
With reordered English corpus	0.20	0.19

Chapter 9

Conclusion and Future Dimension

This thesis attempts to build statistical translation system for English to Urdu and to analyze the errors in the translated text. On the basis of errors analysis, we tried to implement different ways to improve translation. We started with doing experiments with different corpus and found out that performance of the system greatly improved by training on reasonable size parallel corpus.

Since Urdu is morphologically rich language, translation system was unable to generate proper word forms. To handle morphological variations, we built factored translation models with part-of-speech tags and four-lettered stem as factors. We observed that performance of the system improved by aligning on four-lettered stem. But overall, translation quality without factored translation was better. We also did experiment with reordering English sentences before training and it increased the performance of the translation system.

After doing a series of experiments, we can conclude that simpler setup for translation model gives better result for the language pair in consideration. This could be because of the reason that there are not good linguistic tool available for Urdu and accuracy of the tagger was only 70%. Moreover, training data for tagger was also small.

Future Dimensions

In the current situation, translation can be further improved by identifying verb phrase chunk and tagging it. In such way, SOV structure of the sentence can be better learned by the translation system.

Another way to improve the performance could be to add morphological variations of Urdu words in to the corpus. This can be done by implementing a morphological generator which can generate number and gender based inflection of words.

Bibliography

Berger, A. L., Pietra, S. A. D., and Pietra, V. J. D. "A maximum Entropry Approach to Natural Language Processing." *Computational Linguistics*, 1996.

Brown, P.F., S.A.D Pietra, and V.J.D Pietra. "The Mathematics of Statistical Machine Translation: Parameter Estimation." *Association of Computational Linguistics.* 1993.

Callison-Burch, C., Osborne,M.,Koehnn,P. "Re-evaluating the Role of Bleu in Machine Translation Research." European Chapter of Association of Machine Translation, 2006.

Carpaut, M., Wu, D.,. "Improving Statistical Machine Translation using Word Sense Disambiguation." *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning.* Association for Computational Linguistics, 2007.

Chiang, D. "A Hierarchical Phrase-based Model for Statistical Machine Translation." *In Proceedings of ACL.* 2005.

Doddington, G. "Automatic Evaluation of Machine Translation quality using n-gram co-occuring statistics." *In Proceedings of ARPA Workshop on Human language technology.* 2002.

Echizen-ya, H.,Araki,K. "Automatic Evaluation Method for Machine Translation using Noun-Phrase Chunking." *In the Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics.* 2010.

Kirchoff, K.,Rambow,O., Habbash,N., Diab,M. "Semi Automatic Error Analysis for Large-Scale Statistical Machine Translation Systems." *In Proceedings of Machine Translation Summit.* 2007.

Koehn, P., Federico, M., Shen, W., Bertoldi, N., Hoang, H.,Callison-Burch, C., Cowan, B., Zens, R., Dyer, C., Bojar,O., Moran, C., Constantin, A., and Herbst, E. "Open source toolkit for statistical machine translation: Factored translation models and confusion network decoding."

Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico,M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R.,Dyer, C., Bojar, O., Constantin, A., and Herbst, E. "Moses: Open source toolkit for statistical machine translation." *Annual Meeting of the Association for Computational Linguistics.* 2007.

Koehnn, P., and Hieu Hoang. "Factored Translation Models." *Association of Computational Linguistics.* 2007.

Koehnn, P., and K. Knight. "Feature rich translation of Noun Phrases." *41st Annual Meeting of the Association of Computational Linguistics.* 2003.

Koehnn, P., F. J. Och, and D. Marcu. "Statisitcal Phrase based translation." *HLT.* 2003.

Lopez, A. "Statistical Machine Translation." *ACM computing Surveys*, 2008.

Manning, C. D. and Sch¨utze, H. *Foundations of Statistical Natural Language Processing.* MIT Press, 1999.

Murata, M., Uchimoto,K., Ma,Q., Kanamaru,T., Isahara,H. "Analysis of Machine translation Systems' in Tense Aspect and Modality." Proceedings of PACLIC 19, the 19th Asia-Pacific Conference on Language, Information and Computation.

Nießen, S. and Ney, H. "Toward hierarchical models for statistical machine translation of inflected languages." *In the Workshop on Data-Driven Machine Translation at 39th Annual Meeting of the Association of Computational Linguistics.* 2001.

Och, F. J. and Ney, H. "Discriminative training and maximum entropy models for statistical machine translation." *ACL.* 2002.

—. "Improved statistical alignment Models." *ACL.* 2000.

Och, F. J. "Minimum error rate training for statistical machine translation." *Proceedings of the 41st Annual Meeting of Association of Computational Linguistics.* 2003.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. "BLEU: a method for automatic evaluation of machine translation." *ACL.* 2002.

Popovic, M., Gispert,A.D., Gupta,D., Lambert,P., Ney,H., Marino,J.B., Federico,M., Banchs,M. "Morpho-Syntactic information for Automatic Error Analysis of Statistical Machine Translation Output." *Proceedings of the Workshop on Statistical Machine Translation.* Association of Computational Linguistics, 2006.

Sadat, F. and Habash, N. "Combination of arabic preprocessing schemes for statistical machine translation." *Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics.* 2006.

Stolcke. "SRILM - an Extensible Language Modeling Toolkit." *In Proceedings of International Conference on Spoken Language Processing.* 2002.

Vilar, D.,Xu,J., D'Haro,L.F.,Ney,H. "Error Analysis of Machine Translation Output." *In proceedings of LREC.* 2006.