CHAPTER 4

DEVELOPMENT AND IMPROVEMENT OF LANGUAGE MODEL IN TAMIL LANGUAGE

4.1 INTRODUCTION

The strength of a language model lies on factors like capacity, coverage, simplicity and robustness. CIIL has provided Tamil corpus with 761 documents, out of which 500 documents (All sentences in 490 documents and 85% of sentences from 10 documents) with 1.6 million words have been selected. Through manual translation document aligned English corpus for the same has been created which contains 2,491,765 words. Using the Tamil corpus, a baseline language model is created with those 500 Tamil documents. A new method is needed to improve the language model. From the literature, it is understood that improvement of baseline models can be achieved through adaptation by interpolation with translation lexicons generated from English through machine translation or CLSA (Kim and Khudanpur 2003, 2004a, 2004b). English-Tamil CIIL document aligned corpora contain 87,851 unique words in English and 503,567 unique words in Tamil. This large difference in the number of unique words is due to morphological inflections of Tamil. It leads to inaccuracy for both methods. A methodology is needed to bridge the gap between Tamil and English vocabularies and to overcome the probability sparseness of inflected content words of Tamil. These problems are addressed by the application of the proposed partial morphology in Tamil documents before using them in translation techniques.
4.1.1 Generation of Translation Lexicons

Adaptation techniques need unigram, bigram and trigram probabilities of translation lexicons to improve the baseline language model. For creating lexicons, machine translation can be employed directly (Och and Ney 2000) which requires sentence aligned parallel corpora. But it is a bigger constraint for Tamil. Latent semantic analysis is a technique which is essential to capture inherent relationship between term to term (word to word), term to document and document to document. The inherent relationship between words in resource rich and deficient languages is to be identified at the document level since document aligned or story specific corpora are easily available from various sources like television news, newspaper articles, magazines and websites. This can be done through CLSA. After obtaining translation lexicons from CLSA, the baseline language model can be adapted through interpolation with probabilities of translation lexicons. This enables greater improvement to target language model in terms of perplexity and reduction in Word Error Rate (WER) in an application like speech recognition.

4.1.2 Latent Semantic Analysis

LSA is the automatic technique which brings out hidden relations among terms or/and documents in the low dimensional semantic space (Landauer et al 1998). It overcomes problems of pattern matching technique like exact match, non-coverage of synonyms and polysemy (word with different meanings). It is based on the mathematical technique called Singular Value Decomposition (SVD) which finds relations in a low dimensional space with similarity scores between term-term, term-document and document-document (Berry et al 1995). Similarity can be measured by techniques in terms of Spearman correlation ranking and cosine similarity.
4.1.3 Cross-lingual Latent Semantic Analysis

CLSA is the technique which derives the relationship between terms, documents, terms and documents in two different languages. CLSA generates translation lexicons without the translation dictionary and sentence aligned corpora. From the document aligned corpora between two different languages, translation lexicons can be generated when term-term relationship is derived (Dumais et al 1996).

4.1.4 Singular Value Decomposition

SVD is the powerful mathematical technique which factorizes a matrix into three matrices U, S and V as shown in equation (4.1).

\[ A_{m \times n} = U_{m \times m} \times S_{m \times n} \times V_{n \times n}^{T} \]  \hspace{1cm} (4.1)

Here, \( A_{m \times n} \) is a term by document frequency matrix where each document contains source text and its equivalent target language text. \( m \) is the total unique words in both languages (sum of source and target terms) and \( n \) is the total number of documents. \( U_{m \times m} \) is a matrix consisting of columns with eigen vectors of \( A A^{T}_{m \times m} \). \( V_{n \times n} \) is a matrix consisting of columns with eigen vectors of \( A^{T} A_{n \times n} \). \( S_{m \times n} \) is a diagonal matrix consisting of square root of eigen values of \( A A^{T} \) or \( A^{T} A \).

In any language, functional words are well known and identified very easily by automated methods since they occur more frequently across all documents. However content words occur more frequently in relevant documents and less frequently in other documents. When term by document matrix \( A_{m \times n} \) is created, a clear distinction should be made between
functional and content words by using weights. To emphasize the content words, weight factor has to be multiplied with frequency values. This reduces frequency values of functional words and improves the same for content words.

\[ A_{mn} = \begin{pmatrix}
  s_{1,1} & s_{1,2} & \cdots & s_{1,j} & s_{1,j+1} & s_{1,j+2} & \cdots & s_{1,n} \\
  s_{2,1} & s_{2,2} & \cdots & s_{2,j} & s_{2,j+1} & s_{2,j+2} & \cdots & s_{2,n} \\
  \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\
  s_{r,1} & s_{r,2} & \cdots & s_{r,j} & s_{r,j+1} & s_{r,j+2} & \cdots & s_{r,n} \\
  \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\
  s_{sc+1,1} & s_{sc+1,2} & \cdots & s_{sc+1,j} & s_{sc+1,j+1} & s_{sc+1,j+2} & \cdots & s_{sc+1,n} \\
  s_{sc+2,1} & s_{sc+2,2} & \cdots & s_{sc+2,j} & s_{sc+2,j+1} & s_{sc+2,j+2} & \cdots & s_{sc+2,n} \\
  \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\
  t_{1,1} & t_{1,2} & \cdots & t_{1,j} & t_{1,j+1} & t_{1,j+2} & \cdots & t_{1,n} \\
  t_{2,1} & t_{2,2} & \cdots & t_{2,j} & t_{2,j+1} & t_{2,j+2} & \cdots & t_{2,n} \\
  \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\
  t_{m,1} & t_{m,2} & \cdots & t_{m,j} & t_{m,j+1} & t_{m,j+2} & \cdots & t_{m,n}
\end{pmatrix} \]

Let \( sc \) be the total number of source terms, term by document frequency matrix is created with term frequencies \( s_{i,j} \) and \( t_{i,j} \) for source and target language terms respectively (where \( i \) is the index of the term and \( j \) is the index of the document) multiplied by weight \( w_{ij} \) (Husbands et al 2000). Then all the entries in matrix \( A_{mn} \) have term by document frequency with weighted value \( w_{ij} \) as shown in equation (4.2).

\[
 w_{ij} = T_{ij} \cdot \log_2 \frac{n}{n(T_i)} \tag{4.2}
\]

where, \( T_{ij} \) is the \( i^{th} \) term’s frequency in document \( j \), \( n \) is the total number of documents, and \( n(T_i) \) is the number of documents with occurrences of term \( i \).
4.1.5 Reducing LSA Space

To simplify the computation, the dimension of factored matrices can be reduced to $r$ (Berry et al 1995). After dimensionality reduction $A_{m' n}$ becomes $\hat{A}_{m' n}$ as shown in equation (4.3)

$$\hat{A}_{m' n} = U_{m' r} \times S_{r' r} \times V_{r' n}^T$$

(4.3)

This adds strength to computation process by making similar terms to more similar and dissimilar terms to more dissimilar (Kim 2004). $\hat{A}_{m' n}$ matrix and its associated factored matrices $U_{m' r}$, $S_{r' r}$ and $V_{r' n}^T$ serve as trained models to generate translation lexicons for terms in test cases.

4.1.6 Issues in Language Modeling and Adaptation in Tamil Language

Resource deficiency is resolved by means of generating translated lexicons in Tamil from English documents using less expensive CLSA technique. Probability sparseness can be reduced by using Tamil documents in LSA space in which words are processed by partial morphology. CLSA needs trained and reduced LSA space as a base for generation of lexicons from resource rich language and classification of documents with respect to topics or domains. Application of partial morphology in Tamil documents is essential to provide better mapping with English. It avoids sparseness in the similarity score between English and Tamil in the reduced LSA space. It also improves the probability of words based on a lexical root.
4.2 IMPROVEMENT OF LANGUAGE MODEL THROUGH DIRECT ADAPTATION

An initial domain independent model can be developed with any text corpus of available size. This language model can be continuously adapted to improve its performance from English irrespective of the domain or topic. In this experiment, the language model adaptation in domain independent model through document aligned corpora with the application of partial morphology in Tamil documents is proposed.

4.2.1 Partial Morphology in Tamil documents

Morphological analysis is the process of segmenting a given word into a stem and affixes. Preprocessing in terms of partial morphology is needed to minimize the gap in mapping of Tamil and English words. Case endings are to be separated in noun inflected words like (to him) - + and (by Jesus) - + . Other than tense suffix, the remaining verbal suffixes are to be removed in verbal inflections like (saw (subject-masculine)), (saw (subject-feminine)) - , (does (subject-masculine)), (do (subject plural)) - and (will see (subject-masculine)), (will see (subject neutral)) - + . Verb groups are to be separated to map continuous and perfect tenses like (was doing (subject-masculine)) - + and (has done (subject-masculine)) - + . Prefixes which represent determiners are also to be separated like (that boy) - + and (that side) - . Sandhi markers ( , , , ) are to be removed from words. In-depth morphological analysis in Tamil is not needed for CLSA. After partial morphology, unique words in Tamil CIIL corpus have been
reduced to 238,534 words. This has reduced the gap between English and Tamil words significantly. Hence for CLSA, Tamil corpus preprocessed by partial morphology is employed.

### 4.2.2 Training of LSA Space

Bilingual document aligned corpora are used for training LSA space for the generation of translation lexicons. Unique words in English and Tamil for entire corpora are to be created. Term by document frequency matrix $A_{m' n}$ is prepared with terms in rows and documents in columns after finding weighted frequencies of each and every term in documents. Direct mapping can be done for functional words in English-Tamil functional word dictionary since functional words are in closed set and uniformly occur in all documents. For reducing LSA space and minimizing calculations, dimensionality reduction is applied on matrix $A_{m' n}$ by choosing the value $r$ as shown in equation (4.3).

### 4.2.3 Projection of Source Text in the Reduced LSA Space

LSA space is sufficiently trained with bilingual documents. To generate new translation lexicons and probabilities, $p$ number of source documents is taken and matrix $\overline{W}_{m' p}$ is created with weighted frequency values of source words $s_{i, j}$ and 0 assigned to target words:
From $\mathbf{W}$, $\mathbf{S}$ and $\mathbf{V}$ matrices, $\mathbf{U}$ matrix has to be computed for finding target translation lexicons of projected source terms as shown in equation (4.4).

\[
\mathbf{U}_{m \times r} = \mathbf{W}_{m \times p} \cdot \mathbf{V}_{p \times r} \cdot \mathbf{S}_{r \times r}^{-1}
\]  \hspace{1cm} (4.4)

### 4.2.4 Measurement of Similarity between Source and Target Terms

From $\mathbf{U}_{m \times r}$ matrix, cosine similarity values ($\text{Sim}$) are calculated between source and target terms using the expression shown in equation (4.5).

\[
\text{Sim} = \cos \Theta = \frac{\mathbf{X} \cdot \mathbf{Y}}{|\mathbf{X}| \cdot |\mathbf{Y}|}
\]  \hspace{1cm} (4.5)

where $\mathbf{X}$ and $\mathbf{Y}$ are row vectors of source and target terms. From similarity values and their related source terms, correctness is ensured manually for each and every target term and similarity values are used for calculating translation probability using the expression shown in equation (4.6).

\[
P_{\text{CLSA}}(t | s) = \frac{\text{Sim}(t, s)^{g}}{\sum_{i \in T} \text{Sim}(t, s)^{g}} \quad \text{where } g > 1
\]  \hspace{1cm} (4.6)
Here, \( t \) and \( s \) are terms in target and source languages and \( g \) is the power factor which can take a value from 1 to 7. It is seen that power factor improves the perplexity (Coccaro and Daniel Jurafsky 1998).

### 4.2.5 Language Model Adaptation

Unigram probabilities obtained from CLSA are linearly interpolated for terms whose correctness is ensured. For terms which have the same partial stem, the same probability value is used for interpolation. Through this, probabilities of morphologically related content words are boosted due to partial morphology. Language model adaptation is done by linear interpolation by using the formula shown in equation (4.7).

\[
P_{\text{CLSA-INTERPOLATED}}(T_k \mid T_{k-1}, T_{k-2}, S_k) = \lambda P_{\text{CLSA-UNIGRAM}}(T_k \mid S_k) + (1 - \lambda) P(T_k \mid T_{k-1}, T_{k-2}) \quad (4.7)
\]

Here, \( T_k \) and \( S_k \) are target term and its equivalent source term respectively. Optimal interpolation weight \( \lambda \) is obtained by dividing the probability stream into training and test sets. This improves probabilities of content words in domain independent model (Ronald Rosenfeld 1994).

### 4.3 IMPROVEMENT OF DOMAIN INDEPENDENT MODEL THROUGH ADAPTATION WITH TOPIC SPECIFIC MODELS

Adaptation of unigram probabilities obtained through CLSA directly to domain independent language model will yield meagre improvement to content words. To further boost up the probabilities of content words in various topics, translation lexicon probabilities obtained in various topics can be adapted to topic specific models respectively after identification of topics. Later on, the domain independent model can be
adapted with topic specific models through interpolation. This will improve the probabilities of contents words in domain independent model.

4.3.1 Development of Domain Independent and Topic Specific Models

New domain independent model is developed as a base model with available documents in the corpus like CIIL corpus. Multiple topic specific models are developed with their respective topic oriented documents. Topic specific models provide higher accuracy for their respective domain based NLP applications. Domain independent model enables NLP applications in a broader scope across multiple domains. This model will evolve as a vocabulary independent model with greater coverage, accuracy and robustness.

4.3.2 Improvement of Topic Specific and Domain Independent Models

Further adaptations are needed to improve topic specific and domain independent models. This can be done either from documents generated in the domain or from translated lexicons generated from English. It is very difficult to obtain domain information and classify documents pertaining to a domain. To make this happen, automation is needed for classification of documents and also for the generation and improvement of topic specific models. However, improving the domain independent model through direct adaptation with new documents will not yield higher probabilities to content words because they are related to topics. Therefore, topic specific models have to be adapted with new documents which belong to their respective topics. These are identified through topic identification techniques like LSA (Jerome Bellegarda 2000) for the same language and CLSA (Kim 2004) for other language. This will improve probabilities of
content words in their respective domains. After making sufficient improvement in topic specific models, adaptation by interpolation can be employed with domain independent model.

### 4.3.3 Topic Identification of Test Documents

New $\mathbf{\tilde{W}}_{m'}(n+p)$ matrix is created with weighted frequency values of trained $n$ documents and additional $p$ test documents with source terms $s_{i,j}$ upto total number of source terms $(sc)$ and 0 is assigned to target terms.

\[
\begin{pmatrix}
S_{1,1} & S_{1,2} & \cdots & S_{1,n} & S_{1,n+1} & S_{1,n+2} & \cdots & S_{1,n+p} \\
S_{2,1} & S_{2,2} & \cdots & S_{2,n} & S_{2,n+1} & S_{2,n+2} & \cdots & S_{2,n+p} \\
\vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\
S_{n,1} & S_{n,2} & \cdots & S_{n,n} & S_{n,n+1} & S_{n,n+2} & \cdots & S_{n,n+p} \\
0 & 0 & \cdots & 0 & 0 & 0 & \cdots & 0 \\
0 & 0 & \cdots & 0 & 0 & 0 & \cdots & 0 \\
\vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\
0 & 0 & \cdots & 0 & 0 & 0 & \cdots & 0 \\
\end{pmatrix}
\]

In order to identify topics of test documents, $\mathbf{V}^T_{r'}(n+p)$ matrix is calculated using matrices $\mathbf{S}^{-1}_{r',r'}$, $\mathbf{U}^T_{r',m}$ and $\mathbf{\tilde{W}}_{m'}(n+p)$ as shown in equation (4.8)

\[
\mathbf{V}^T_{r'(n+p)} = \mathbf{S}^{-1}_{r',r'} \times \mathbf{U}^T_{r',m} \times \mathbf{\tilde{W}}_{m'(n+p)} \tag{4.8}
\]

### 4.3.4 Measurement of Similarity between Test Documents and Topics

From matrix $\mathbf{V}^T_{r'(n+p)}$, cosine similarity values are calculated between test documents and topic based trained documents using the
expression shown in equation (4.5). Here, \( \mathbf{X} \) and \( \mathbf{Y} \) are column vectors of test
document and topic based trained document respectively. From the similarity
values, topics of \( p \) test documents are identified.

4.3.5 Adaptation in Topic Specific Language Models

Unigram probabilities of a test document obtained from CLSA
whose correctness is ensured will be linearly interpolated in the respective
topic specific model after identification of the topic. Language model
adaptation by linear interpolation (Ronald Rosenfeld 1994) is done using the
formula shown in equation (4.7). Here, \( T_k \) and \( S_k \) are target term and its
equivalent source term respectively. This improves probabilities of content
words in topic specific models.

4.3.6 Adaptation in Domain Independent Model

After sufficient adaptations of topic specific models, the domain
independent model is adapted with topic specific models through
interpolation technique as shown in equation (4.9).

\[
P_{\text{Domain-Ind-Adapted}} = \sum_{i \in 2^{\text{no of topics}}} \alpha_i \left( P_{\text{Domain-Ind}} + \alpha \sum_{i \in 2^{\text{no of topics}}} \beta_i \right) P_{\text{Topic}\_i}
\]  

(4.9)

where \( i \) is the topic index. This domain independent model will have greater
coverage, accuracy and robustness (Ronald Rosenfeld 1994).

4.4 EXPERIMENTS

Initial domain independent model has been developed using
statistical language modeling toolkit of Carnegie Melon University (CMU)
with 500 CIIL Tamil documents (All sentences in 490 documents and 85% of
sentences from 10 documents) which contain 1,627,150 words. A test set
which contains 607 sentences (4042 words) is used for the calculation of perplexity.

### 4.4.1 Creation of Reduced LSA Space

By using Tamil-English document aligned corpora with the same 500 documents, reduced LSA space has been created using CLSA. After partial morphology, the total number of words in Tamil documents has been increased to 2,430,034 words and the number of unique words has been reduced to 238,534 words. This reduces the gap between English and Tamil words in LSA space and improves translation accuracy. Term by document matrix $A_{326385 \times 500}$ has been created (87,851 +238,534 = 326,385 terms and 500 documents) with a weight factor. Dimensionality reduction has also been done with $r$ value as 100 and $U_{326385 \times 100}$, $S_{100 \times 100}$, and $V^{T}_{100 \times 500}$ matrices have been created as reduced LSA space.

### 4.4.2 Direct Adaptation in Domain Independent Model

In the first experiment, direct adaptation is done in two stages. In the first stage, 5 additional documents with 2587 words in English (1582 unique words) are taken and $\tilde{W}_{326385 \times 5}$ matrix is created with weighted frequency values for English terms and 0 for Tamil terms. From the matrices $\tilde{W}_{326385 \times 5}$, $S_{100 \times 100}$, and $V^{T}_{5 \times 100}$, $U_{326385 \times 100}$ matrix is obtained using equation (4.4). Using cosine similarity, for all the unique English terms occurring in projected documents, their equivalent Tamil terms and their similarity values are obtained and correctness is ensured manually. From similarity values, translation probabilities are calculated and interpolated in the domain independent model. In the second stage, another
adaptation is done with another 5 documents which contain 2905 words (1751 unique words). The details are shown in Table 4.1.

### Table 4.1 Details of CLSA and Adaptations

<table>
<thead>
<tr>
<th>Details</th>
<th>CLSA</th>
<th>First Adaptation</th>
<th>Second Adaptation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Corpora</td>
<td>CIIL English-Tamil Document aligned Corpora</td>
<td>English corpus</td>
<td>English corpus</td>
</tr>
<tr>
<td>No. of Documents</td>
<td>500</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>No. of words</td>
<td>2,491,765 (English)</td>
<td>2,430,034 (Tamil - after partial morphology)</td>
<td>2587</td>
</tr>
<tr>
<td>No. of unique words</td>
<td>87,851 (English)</td>
<td>238,534 (Tamil – after partial morphology)</td>
<td>1582</td>
</tr>
<tr>
<td>Translation accuracy in</td>
<td>NA</td>
<td>70 %</td>
<td>72 %</td>
</tr>
<tr>
<td>CLSA</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

NA – Not Applicable

### 4.4.3 Adaptation in Domain Independent Model with Topic Specific Models

In the second experiment, 85%, 10% and 5% sentences from those 10 documents are used for development, adaptation and testing of topic specific models respectively and 10 topic specific models are created. Using 10 documents created with 10% sentences of selected topics, \( \mathbf{W}_{326385 \times 10} \) matrix is created with weighted English term frequencies and 0 is assigned to Tamil term frequencies. From \( \mathbf{W}_{326385 \times 10} \mathbf{S}^{-1} \mathbf{1}_{100} \mathbf{1}^{T} \) and
Using cosine similarity, for all unique English terms occurring in projected 10 documents, their equivalent Tamil terms and similarity values are obtained and correctness is ensured manually. From similarity values, Tamil lexicon probabilities are obtained and stored as probability streams.

For topic identification, a new $\mathbf{W}_{326385 \times 510}$ matrix is created using the same 500 documents and 10 documents which contain 10% of sentences of selected topics with weighted English term frequencies and 0 is assigned to Tamil term frequencies. From $\mathbf{W}_{326385 \times 510}$, $\mathbf{S}_{100 \times 100}$ and $\mathbf{U}_{326385 \times 100}$ matrices, $\mathbf{V}_{100 \times 510}$ matrix is obtained using equation (4.8). Using cosine similarity, topics are identified for the selected 10 documents. After identification of topic, probability streams obtained from $\mathbf{U}_{326385 \times 100}$ matrix are adapted into their respective topic specific models. Thus a domain independent model is finally adapted with topic specific models.

### 4.5 RESULTS AND DISCUSSION

After each direct adaptation in the domain independent language model, the perplexity of the language model is tested with the same test set. This language model is used in an in-house Automatic Speech Recognizer (ASR) and WER is obtained for each stage. The results are summarized in Table 4.2.
Table 4.2 Results of Domain Independent Model through Direct Adaptations

<table>
<thead>
<tr>
<th>Details</th>
<th>Domain independent Model</th>
<th>First Adaptation</th>
<th>Second Adaptation</th>
</tr>
</thead>
<tbody>
<tr>
<td>No. of documents</td>
<td>500 (Tamil)</td>
<td>5 (English)</td>
<td>5 (English)</td>
</tr>
<tr>
<td>No. of words</td>
<td>1,627,150 (Tamil)</td>
<td>2587 (English)</td>
<td>2905 (English)</td>
</tr>
<tr>
<td>Sentences in test case</td>
<td>607 (12 documents)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Words in test case</td>
<td></td>
<td>4042</td>
<td></td>
</tr>
<tr>
<td>Perplexity</td>
<td>121.71</td>
<td>119.94</td>
<td>118.04</td>
</tr>
<tr>
<td>Word error rate</td>
<td>7.94 %</td>
<td>7.89%</td>
<td>7.86%</td>
</tr>
</tbody>
</table>

A comparison of perplexity and WER after direct adaptation in domain independent language model is set forth in Figure 4.1 and 4.2 respectively.

![Figure 4.1 Comparison of Perplexity in Direct Adaptation](image)
Figure 4.2 Comparison of WER in Direct Adaptation

10 topic specific models are created with 85% sentences from 10 documents and perplexity values are obtained with test sets which contain 5% sentences from the same documents. After adaptation with 10% sentences from the same documents, perplexity values are obtained using the same test sets. The details are listed in Table 4.3. This discussion presented so far demonstrates clearly the reasonable improvements are obtained after adaptations.

Table 4.3 Details of Topic Specific Models and their Perplexity Values

<table>
<thead>
<tr>
<th>Topic No.</th>
<th>Total Sentences</th>
<th>Training</th>
<th>Adaptation</th>
<th>Testing</th>
<th>Perplexity</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Sentences</td>
<td>Words</td>
<td>Sentences</td>
<td>Words</td>
</tr>
<tr>
<td>1</td>
<td>496</td>
<td>422</td>
<td>4028</td>
<td>50</td>
<td>474</td>
</tr>
<tr>
<td>2</td>
<td>505</td>
<td>429</td>
<td>4173</td>
<td>51</td>
<td>516</td>
</tr>
<tr>
<td>3</td>
<td>659</td>
<td>561</td>
<td>4234</td>
<td>66</td>
<td>615</td>
</tr>
<tr>
<td>4</td>
<td>522</td>
<td>442</td>
<td>4143</td>
<td>52</td>
<td>515</td>
</tr>
<tr>
<td>5</td>
<td>506</td>
<td>430</td>
<td>4153</td>
<td>51</td>
<td>524</td>
</tr>
<tr>
<td>6</td>
<td>527</td>
<td>448</td>
<td>4069</td>
<td>53</td>
<td>423</td>
</tr>
<tr>
<td>7</td>
<td>563</td>
<td>476</td>
<td>4107</td>
<td>56</td>
<td>420</td>
</tr>
<tr>
<td>8</td>
<td>571</td>
<td>485</td>
<td>3783</td>
<td>57</td>
<td>473</td>
</tr>
<tr>
<td>9</td>
<td>537</td>
<td>457</td>
<td>4076</td>
<td>54</td>
<td>439</td>
</tr>
<tr>
<td>10</td>
<td>565</td>
<td>480</td>
<td>3682</td>
<td>57</td>
<td>489</td>
</tr>
</tbody>
</table>
Similarly, the perplexity of domain independent model is calculated before and after adaptation through topic specific models. This results in a significant improvement of domain independent model. This domain independent model is used in ASR and WER is obtained before and after adaptation. The results are shown in Table 4.4.

Table 4.4 Results of Domain Independent Model adapted with Topic Specific Models

<table>
<thead>
<tr>
<th>Details</th>
<th>Perplexity</th>
<th>WER in ASR</th>
</tr>
</thead>
<tbody>
<tr>
<td>Before adaptation</td>
<td>121.71</td>
<td>7.94</td>
</tr>
<tr>
<td>After adaptation</td>
<td>115.56</td>
<td>7.81</td>
</tr>
</tbody>
</table>

A comparison of perplexity and WER before and after adaptation of topic specific models in domain independent model is displayed in Figure 4.3 and 4.4 respectively.

![Figure 4.3 Comparison of Perplexity in adaptation through Topic Specific Models](image-url)
In the same way, a comparison of perplexity and WER before adaptation and after direct and topic specific model adaptations in domain independent model is shown in Figure 4.5 and 4.6 respectively.

Figure 4.5  Comparison of Perplexity in Direct and Topic Specific Models Adaptation
Figure 4.6  Comparison of WER in Direct and Topic Specific Models Adaptation

4.6 SUMMARY

In this work, initially a domain independent model has been created and directly adapted in two stages. Secondly, a new domain independent and 10 topic specific models have been created. Then, the domain independent model has been adapted with topic specific models. The results show significant improvements in perplexity and WER, obtained before and after direct adaptation with domain independent model. The probabilities of content words have been further boosted through topic specific models rather than through direct adaptation.