Language model adaptation in Tamil language using cross-lingual latent semantic analysis with document aligned corpora

M. Selvam[1,*] and A. M. Natarajan[2]

[1]Department of Information Technology, Kongu Engineering College Perundurai, Erode 638 052, India
[2]Department of Electronics and Communication Engineering, Bannari Amman Institute of Technology, Sathyamangalam, Erode 638 401, India

Unlike English, Tamil does not have sufficient volume of text corpus to build a reliable language model. In this work, domain independent language model has been built with 500 Tamil documents. To improve the language model, adaptation with translation lexicons in Tamil generated from English using cross-lingual latent semantic analysis (CLSA) has been employed. Since Tamil is an agglutinative language, usage of surface word forms in CLSA will not yield better translation accuracy. Lexical gap between English and Tamil words has been reduced by the proposed partial morphological analysis in Tamil. This has improved the translation accuracy. Experiments have been conducted with direct and topic-specific model adaptations to improve the domain independent model. Significant improvements have been obtained in terms of perplexity and word error rate.

Keywords: Adaptation, cross-lingual latent semantic analysis, document aligned corpora, language model, Tamil language.

LANGUAGE model is the probability distribution over a sequence of strings. It is gaining momentum in the field of natural language processing (NLP). Applications like speech recognition, machine translation, optical character recognition, hand-written character recognition, spelling correction, document classification and information retrieval need better language models[1]. A language model is the mechanism which assists in generating output sentences from the sequence of words generated by any NLP application. Some examples include *n*-gram language model, distance-based language model and class-based language model[2].

The output of any NLP application is the optimal sequence of words which are available in the dictionary. This can be achieved by means of a language model. Pure grammar-driven models strictly restrict the sequence of words and more computational efforts are needed to make the system recognize the sequence. Grammar-based methods focus on syntax with a long distance relationship among words, manifested in parse trees which are widely used for small vocabulary tasks. Statistical methods are primarily data driven. The frequencies of patterns as they occur in any training corpora are recorded as the probability distributions. These methods, with *n*-gram or trigram approach, mainly focus on short distance relationship among words in sentences which depend on a large training set and they are suitable to model large vocabulary tasks.

A language model needs a large volume of training corpus for coverage, accuracy and robustness. Languages like English and French have sufficient resources like speech and text corpora. But, the development and improvement of language model in Tamil is challenging due to two major factors like resource deficiency and morphological richness. Tamil is highly resource deficient in terms of preprocessed text corpora and it has longer morphological inflections in its words. Hence, development and improvement of a language model with surface level word forms become less efficient even for a corpus of reasonable size due to the uniqueness of words. In fact, they involve many issues.

Issues in development of language model in Tamil language

To begin with, two major issues in the development of language model in Tamil are resource deficiency, in terms of preprocessed text corpora, and probability sparseness. Resource deficiency does not allow a language model to provide greater coverage, accuracy and robustness in an application. Probability sparseness refers to the poor probability value of a word due to longer agglutination (combination of many morphemes). For a root word, there may be many word forms with different affixes. The probability value of a root word gets diffused among its inflectional word forms. When surface forms of Tamil words are used in statistical language modelling, poor probability values are obtained for content words even as a large volume of corpus is used to train the model.

*For correspondence. (e-mail: amm_selvam@yahoo.co.in)

Issues in improvement of language model in Tamil language

Improvement of language model in Chinese was successfully carried out through adaptation with translation lexicons in Chinese, generated from English. Translation lexicons in Chinese were obtained through machine translation with English–Chinese sentence aligned corpora[3] and cross-lingual latent semantic analysis (CLSA) with English–Chinese document aligned corpora[4–7]. A language model adaptation in domain independent Chinese language model was done with direct and topic specific models adaptations. English was preferred due to its richness in resources, tools and techniques[5,6].

Similarly, a language model in Tamil can be created with the available corpus and it needs to be continuously adapted with translation lexicons in Tamil, generated from English through machine translation or CLSA for its improvement. Yet, the translation accuracy will be poor due to a lexical gap between Tamil and English words. 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. From the remaining 15% of sentences of 10 documents, 10% sentences are used for adaptation and 5% sentences are used for testing in the experiments. Through manual translation, a document aligned English corpus, for the same 500 documents has been created. It contains 2.5 million words. In this English–Tamil CIIL document aligned corpora, there are 87,851 and 503,567 unique words in English and Tamil respectively. This large difference in the number of unique words is due to morphological inflections of Tamil. It leads to inaccuracy for both machine translation and CLSA methods.

Need for partial morphological analysis

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 CLSA. In this article, the aforesaid issues are addressed in the development and improvement of language models in Tamil. In order to improve the performance of language models, various experiments are conducted with direct and topic specific model adaptations. The performance is evaluated in terms of perplexity using Carnegie Mellon University Statistical Language Modeling (CMU SLM) toolkit and word error rate (WER) using in-house automatic speech recognizer (ASR).

Background

Adaptation techniques need unigram probabilities of lexicons to improve a baseline n-gram model. For creating lexicons directly, the machine translation can be employed using sentence aligned corpora[5]. But sentence aligned corpora are rarely available. This is a constraint for resource deficient languages. Latent semantic analysis (LSA) is a technique used to capture inherent relationship between term to term (word to word), term to document and document to document[8]. It is based on a mathematical technique called singular value decomposition (SVD) which finds relations in the low-dimensional space with similarity scores. This similarity can be measured by techniques like Spearman correlation ranking and cosine similarity. In the experiment, cosine similarity is used due to its shorter running time than Spearman correlation ranking. The inherent relationship between words in resource rich and deficient languages is to be identified at the document level[9] since document aligned or story specific corpora are easily available from various sources like television news, newspaper articles, magazines and websites. CLSA can be employed for this purpose. Adaptation with translation probabilities enables greater improvement to a language model in terms of perplexity and reduction in WER in applications like speech recognition and machine translation.

Singular value decomposition

SVD is a powerful mathematical technique[10] which factorizes any $m \times n$ matrix into three matrices $\mathbf{U}$, $\mathbf{S}$ and $\mathbf{V}$ as shown in eq. (1).

$$\mathbf{A}_{m\times n} = \mathbf{U}_{m\times m} \times \mathbf{S}_{m\times n} \times \mathbf{V}^{\mathbf{T}}_{n\times n} \tag{1}$$

Here, $\mathbf{A}_{m\times n}$ is term by document frequency matrix, where each document contains a source text and its equivalent text in target language, m the total unique words in both languages and n the total number of documents, $\mathbf{U}_{m\times m}$ a matrix consisting of columns with eigenvectors of $\mathbf{AA}^{\mathbf{T}}$, $\mathbf{S}_{m\times n}$ a diagonal matrix consisting of square root of eigenvalues of $\mathbf{AA}^{\mathbf{T}}$ or $\mathbf{A}^{\mathbf{T}}\mathbf{A}$, $\mathbf{V}_{m\times n}$ a matrix consisting of columns with eigenvectors of $\mathbf{A}^{\mathbf{T}}\mathbf{A}$.

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 a term by document matrix $\mathbf{A}_{m\times n}$ is created, a clear distinction should be made between functional and content words by using weights[11]. Entries in matrix $\mathbf{A}_{m\times n}$ are term by document frequencies, with weighted values W_{ij} as shown in eq. (2).

$$W_{ij} = T_{ij} \times \log_2 \frac{n}{n(t_i)}, \tag{2}$$

where T_{ij} is the ith term's frequency in document j, n the total number of documents and $n(t_i)$ the number of documents with occurrences of term i.

Reducing LSA space

To simplify computation, dimensions of factored matrices can be reduced[10] to r. Then eq. (1) becomes

$$\hat{\mathbf{A}}_{m\times n} = \mathbf{U}_{m\times r} \times \mathbf{S}_{r\times r} \times \mathbf{V}^{\mathbf{T}}_{r\times n}. \quad (3)$$

Dimensionality reduction adds strength to computation process by making similar terms to more similar and dissimilar terms to more dissimilar[12]. Matrix $\hat{\mathbf{A}}_{m\times n}$ and its associated factored matrices $\mathbf{U}_{m\times r}, \mathbf{S}_{r\times r}, \mathbf{V}^{\mathbf{T}}_{r\times n}$ will serve as trained models to generate translation lexicons for terms in test cases which will be used for language model adaptation.

Nature of Tamil language

The grammar of Tamil is agglutinative in nature. Suffixes are used to mark class, number and cases attached to a noun. Words consist of a lexical root to which one or more affixes are attached. Most of the affixes are suffixes which can be derivational or inflectional. The length and extent of agglutination are longer which result in longer words with a large number of suffixes. For example, a Tamil word can contain a stem with one prefix and seven suffixes at the maximum. Some of the other issues are morpho-phonology (*sandhi*) rules, complex noun and verbal patterns, and out of vocabulary rate due to inflections. Poetic forms are more complex than prose forms.

In Tamil, nouns are classified into rational and irrational forms. Humans come under rational form whereas all other nouns are classified as irrational. Rational nouns and pronouns belong to one of the three classes: masculine singular, feminine singular and rational plural. Irrational nouns belong to one of the two classes: irrational singular and irrational plural. Morphological inflections on nouns include gender and number. Suffixes are used to perform functions of cases or postpositions with nouns like ablative, accusative, benefactive, clitics, dative, genitive, instrumental, locative, oblique, selective and sociative[13]. Verbs are also inflected through various suffixes which indicate person, number, gender, mood, tense, honorific and voice[14]. Some other forms of verbs are transitive, intransitive, causative, infinitive, imperative and reportive. Adjective comes along with tense or negative participles. Prepositions take direct and noun inflected forms. Other parts of speech take simpler forms.

Tamil is consistently head-final language. Verb comes at the end of the clause with a typical word order of subject object verb (SOV). However, Tamil allows word order to be changed, making it a relatively word order free language. Other features are plural for honorific noun, frequent echo words, and null subject feature, i.e. some sentences do not have subject, verb and object.

Significance of CLSA and partial morphological analysis

Resource deficiency is resolved by means of generating translated lexicons in Tamil from English documents using a less expensive CLSA technique. Probability sparseness can be reduced by using Tamil documents in LSA space in which words are processed by partial morphological analysis. CLSA needs a trained and reduced LSA space as a base for the generation of lexicons from English and classification of documents with respect to topics or domains. Partial morphological analysis in Tamil is essential to provide better mapping, avoid sparseness in the similarity score between English and Tamil in the reduced LSA space, and improve the probability of the words based on the lexical root.

Partial morphological analysis

Morphological analysis is the process of segmenting a given word into a stem and its affixes. The following preprocesses in terms of partial morphology are needed to minimize the gap in mapping of Tamil and English words.

a. Case endings are to be separated in noun inflected words. The stem and suffixes are to be used separately. For example

அவனுக்கு (to him) → அவன் + க்கு
இயேசுவால் (by Jesus) → இயேசு + ஆல்.

b. Other than tense suffix, remaining verbal suffixes are to be removed in verbal inflections. Partial stem attached with present or past tense suffix is used in CLSA. For the future tense verb, stem and tense suffix are separately used. For example

பார்த்தான் (saw) → பார்த்
(Verb inflected with third person, singular, male, simple past tense)

பார்த்தாள் (saw) → பார்த்
(Verb inflected with third person, singular, female, simple past tense)

செய்கிறான் (does) → செய்கிறு
(Verb inflected with third person, singular, male, simple present tense)

செய்கிறார்கள் (do) → செய்கிறு
(Verb inflected with third person, plural, neutral, simple present tense)

பார்ப்பான் (will see) → பார் + ப்
(Verb inflected with third person, singular, male, simple future tense)

பார்க்கும் (will see) → பார் + ப்
(Verb inflected with third person, singular, neutral, simple future tense)

செய்வான் (will do) → செய் + வ்
(Verb inflected with third person, singular, male, simple future tense)

c. Verb groups are to be separated and used to map continuous and perfect tenses. For example,

செய்துகொண்டிருந்தான் (was doing) → செய்து + கொண்டிருந்தான்
(Verb inflected with third person, singular, male, past continuous tense)

செய்துவிட்டான் (has done) → செய்து + விட்டான்
(Verb inflected with third person, singular, male, present perfect tense)

d. Prefixes which represent determiners are also to be separated. Prefix and stem are used for mapping. For example,

அந்தப்பையன் (the boy) → அந்த + பையன்
அப்பக்கம் (the page) → அ + பக்கம்

e. *Sandhi* markers (க், ச், த், ப்) are to be removed from the words. Words without *sandhi* markers are used for the mapping.

In-depth morphological analysis in Tamil is not necessary for this CLSA. After partial morphology, the number of unique words in Tamil corpus is reduced to 238,534 words. This reduces the gap between English and Tamil words significantly. Here, preprocessed Tamil corpus by partial morphology is employed for CLSA.

Development and improvement of language models

A domain independent language model in Tamil is created with available documents. It serves as a common model for multiple domains. Topic specific language models in Tamil are also created and used for domain-specific applications. In order to improve domain independent and topic specific models in Tamil, lexicon generation from English, document classification and adaptation are essential. After building the reduced LSA space, lexicons and their probabilities in Tamil are generated from new English documents projected in the reduced LSA space. Translation probabilities can be adapted to the domain independent language model directly or topic-specific models which in turn adapted to the domain independent model. Adaptation to topic-specific models needs classification of documents or identification of topics through the same LSA space.

Improvement of language model with 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. In this section, the language model adaptation by interpolation with translation lexicon probabilities in Tamil from English using CLSA is proposed.

Training of LSA space: English–Tamil document aligned corpora are used for training LSA space. Unique words in both languages for the entire corpora are to be created. Term by document frequency matrix is prepared with terms in rows and documents in columns after finding frequencies of each and every term in the respective documents. To emphasize the content words, the weight factor has to be multiplied with frequency values. This reduces frequency values of functional words and improves the same for content words. Direct mapping can be done for functional words in English–Tamil functional word dictionary since functional words are in closed set and they uniformly occur in all documents.

Let $\mathbf{A}_{m\times n}$ be the term by document frequency matrix which is decomposed into three matrices, namely $\mathbf{U}_{m\times m}$, $\mathbf{S}_{m\times n}$ and $\mathbf{V}^{\mathbf{T}}_{n\times n}$ using SVD as shown eq. (1). For reducing the LSA space and minimizing the calculations, dimensionality reduction is applied by choosing the value r as shown in eq. (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 $\bar{\mathbf{W}}$ is created with weighted frequency values of source words S_{ij} and 0 for target words.

$$\bar{\mathbf{W}} = \begin{bmatrix} s_{1,1} & s_{1,2} & \cdots & s_{1,p} \\ s_{2,1} & s_{2,2} & \cdots & s_{2,p} \\ \vdots & \vdots & \ddots & \vdots \\ s_{sc,1} & s_{sc,2} & \cdots & s_{sc,p} \\ 0 & 0 & \cdots & 0 \\ 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 0 \end{bmatrix}$$

From $\bar{\mathbf{W}}$, $\mathbf{S}$ and $\mathbf{V}$ matrices, $\mathbf{U}$ matrix has to be computed for finding target translation lexicons of the projected source terms as shown in eq. (4):

$$\mathbf{U}_{m\times r} = \bar{\mathbf{W}}_{m\times p} \times \mathbf{V}_{p\times r} \times \mathbf{S}^{-1}_{r\times r}. \quad (4)$$

Measurement of similarity between source and target terms: From $\mathbf{U}_{m\times r}$ matrix, the cosine similarity values are calculated between source and target terms using the expression shown in eq. (5).

$$\text{Similarity value} = \cos\theta = \frac{\mathbf{A}\cdot\mathbf{B}}{|\mathbf{A}|\cdot|\mathbf{B}|} \quad (5)$$

Here, **A** and **B** are the 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 eq. (6).

$$P_{\text{CLSA}}(t\,|\,s) = \frac{\text{Sim}(t,s)^{\gamma}}{\sum_{t'\in T}\text{Sim}(t,s)^{\gamma}}, \quad \text{where } \gamma >> 1. \quad (6)$$

Here t and s are terms in target and source languages and γ the power factor which can take a value from 1 to 7. It is seen that power factor improves the perplexity[15].

Language model adaptation: Unigram probabilities obtained from CLSA will be linearly interpolated for terms whose correctness has been ensured. For terms which have the same partial stem, the same probability value will be used for interpolation. Through this, probabilities of morphologically related content words will be boosted due to partial morphology. Language model adaptation by linear interpolation is done by using the formula shown in eq. (7).

$$P_{\text{CLSA-INTERPOLATED}}(T_k\,|\,T_{k-1},T_{k-2},S_k) = \lambda P_{\text{CLSA-Unigram}}(T_k\,|\,S_k) + (1-\lambda)P(T_k\,|\,T_{k-1},T_{k-2}). \quad (7)$$

Here, T_k and S_k are target and its equivalent source term. Optimal interpolation weight λ is obtained by dividing the probability stream into training and testing set[16]. This improves the probability of content words in the domain independent model.

Improvement of domain independent model with topic-specific models

The adaptation of unigram probabilities obtained through CLSA directly to domain independent language model yields only meagre improvement to content words. To further boost 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, the domain independent model can be adapted with topic specific models through interpolation. This will further improve the probabilities of contents words in the domain independent model.

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

Improvement of topic-specific and domain independent models: Further adaptations are needed to improve topic specific and domain independent models. Obtaining domain information or classifying documents pertaining to a domain is normally very difficult. Automation is needed for classification of documents and also for generation and improvement of topic specific models. However, improvement of the domain independent model through direct adaptation with new documents will not yield high probabilities to content words because they are related to topics. Therefore, topic specific models have to be adapted with new documents after the identification of topics through techniques like LSA[17] for the same language and CLSA[12] for the other language. This will improve the probabilities of content words in their respective domains. After making sufficient improvement in topic specific models, adaptation by interpolation can be employed with the domain independent model.

Topic identification of test documents: New $\bar{\mathbf{W}}_{m\times(n+p)}$ matrix is created with weighted frequency values of trained n documents and additional p test documents with source terms $S_{i,j}$ up to the total number of source terms sc and 0 assigned to target terms.

$$\bar{\mathbf{W}} = \begin{bmatrix} s_{1,1} & s_{1,2} & \cdots & s_{1,n} & s_{1,n+1} & s_{1,n+2} & \cdots & s_{1,n+p} \\ s_{2,1} & s_{2,2} & \cdots & s_{2,n} & s_{2,n+1} & s_{2,n+2} & \cdots & s_{2,n+p} \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\ s_{sc,1} & s_{sc,2} & \cdots & s_{sc,n} & s_{sc,n+1} & s_{sc,n+2} & \cdots & s_{sc,n+p} \\ 0 & 0 & \ldots & 0 & 0 & 0 & \ldots & 0 \\ 0 & 0 & \ldots & 0 & 0 & 0 & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & 0 & 0 & 0 & \ldots & 0 \end{bmatrix}.$$

In order to identify the topic of test documents, $\mathbf{V}^{\mathbf{T}}_{r\times(n+p)}$ matrix is calculated using matrices, $\mathbf{S}^{-1}_{r\times r}$, $\mathbf{U}^{\mathbf{T}}_{r\times m}$ and $\hat{\mathbf{W}}_{m\times(n+p)}$ as shown in eq. (8):

$$\mathbf{V}^{\mathbf{T}}_{r\times(n+p)} = \mathbf{S}^{-1}_{r\times r} \times \mathbf{U}^{\mathbf{T}}_{r\times m} \times \hat{\mathbf{W}}_{m\times(n+p)}. \quad (8)$$

Measurement of similarity between test documents and topics: From matrix $\mathbf{V}^{\mathbf{T}}_{r\times(n+p)}$, cosine similarity values are calculated between test documents and topic-based trained documents, using the expression shown in eq. (5). Here **A**, **B** are column vectors of test document and topic-based trained documents respectively. From similarity values, topics of p test documents are identified.

Adaptation in topic-specific language models: Unigram probabilities of a test document, obtained from CLSA whose correctness has been ensured will be linearly interpolated in the respective topic-specific model after identification of topics. In this experiment, 10 topic-specific models are created. Language model adaptation by linear interpolation[16] is done using the formula shown in eq. (7). Here, T_k and S_k are target and its equivalent source term. This improves the probabilities of content words in topic specific models.

Adaptation in domain independent model: After sufficient adaptations of topic-specific models, the domain independent model has to be adapted with the topic-specific models through interpolation technique[16] as shown in eq. (9):

$$P_{\text{Domain-Ind-Adapted}} = \lambda_1 \times P_{\text{Domain-Ind}} + \sum_{i=2 \text{ to } N} \lambda_i \times P_{\text{Topic}-i}, \quad (9)$$

where i is the topic index varying from 2 to 11 (10 topics). This domain independent model will have greater coverage, accuracy and robustness.

Experiments

In the experiments, an initial domain independent model is developed using CMU SLM toolkit with 500 CIIL Tamil documents (all sentences in 490 documents and 85% of sentences from 10 documents) which contain 1,627,150 words. Ten per cent sentences of 10 documents are used for two stages of adaptation each with five documents. Five per cent sentences of those 10 documents are used for testing perplexity values in topic-specific models. Another test set which comprises 607 sentences (4042 words) is used for the calculation of perplexity in the domain independent language model before and after adaptations. An in-house partial morphological analyser has been built with an accuracy of 93% and used for segmenting Tamil words to map with English words.

Creation of reduced LSA space

By using document aligned corpora with the same 500 documents in Tamil and their equivalent in-house translated 500 English documents which contain 2,491,765 words, reduced LSA space is created using CLSA. The document aligned corpora contain 87,851 unique words in English and 503,567 unique words in Tamil. After partial morphology, the total words in Tamil documents are increased to 2,430,034 words and their unique words are 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 $\mathbf{A}_{326385\times500}$ is created (87,851 + 238,534 = 326,385 terms and 500 documents) with weighting factor. Dimensionality reduction is done with r value as 100 and $\mathbf{U}_{326385\times100}$, $\mathbf{S}_{100\times100}$, $\mathbf{V}^{\mathbf{T}}_{100\times500}$ matrices are created as reduced LSA space.

Direct adaptation in domain independent model

Adaptation is done in two stages. In the first stage, five additional documents with 2587 words in English (1582 unique words) are taken and $\bar{\mathbf{W}}_{326385\times5}$ matrix is created with weighted frequency values for English terms and 0 for Tamil terms. From the matrices $\bar{\mathbf{W}}_{326385\times5}$, $\mathbf{S}^{-1}_{100\times100}$ and $\mathbf{V}_{5\times100}$, **U** matrix is obtained using eq. (4). Using cosine similarity for all unique English terms occurring in projected documents, 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 five documents which contain 2905 words (1751 unique words). The details are shown in Table 1.

Adaptation in domain independent model with topic-specific models

For development, adaptation and testing of topic-specific models, 85%, 10% and 5% sentences from those 10 documents are used respectively and 10 topic-specific models are created. Using 10 documents created with 10% sentences of selected topics which have to be used for adaptation, $\bar{\mathbf{W}}_{326385\times10}$ matrix is created with weighted English term frequencies and 0 for Tamil term frequencies. From $\bar{\mathbf{W}}_{326385\times10}$, $\mathbf{S}^{-1}_{100\times100}$ and $\mathbf{V}_{10\times100}$ matrices, $\mathbf{U}_{326385\times100}$ matrix is obtained using eq. (4). Using cosine similarity, for all unique English terms occurring in projected 10 documents, equivalent Tamil terms and their similarity values are obtained and correctness is ensured manually. From similarity values, Tamil lexicon probabilities are obtained and stored as probability streams.

Table 1. Details of CLSA and adaptations

Details	CLSA	First adaptation	Second adaptation
Corpora	CIIL English–Tamil document aligned Corpora	English corpus	English corpus
No. of documents	500	5	5
No. of words	2,491,765 (English) 1,627,150 (Tamil) 2,430,034 (Tamil – after partial morphology)	2587	2905
No. of unique words	87,851 (English) 503,567 (Tamil) 238,534 (Tamil – after partial morphology)	1582	1751
Translation accuracy	NA	70%	72%

NA, Not applicable.

Table 2. Results of domain independent model with direct adaptations

Details	Domain independent model	First adaptation	Second adaptation
No. of documents	500 (Tamil)	5 (English)	5 (English)
No. of words	1,627,150 (Tamil)	2587 (English)	2905 (English)
Sentences in test case	607 (12 documents)		
Words in test case	4042		
Perplexity	121.71	119.94	118.04
Word error rate	7.94%	7.89%	7.86%

Table 3. Details of topic-specific models and their perplexity values

		Training		Adaptation		Testing		Perplexity	
Topic no.	Total sentences	Sentences	Words	Sentences	Words	Sentences	Words	Initial	After adaptation
1	496	422	4028	50	474	24	214	772.94	756.422
2	505	429	4173	51	516	25	217	899.90	799.142
3	659	561	4234	66	615	32	283	963.10	925.888
4	522	442	4143	52	515	28	279	628.42	496.372
5	506	430	4153	51	524	25	236	989.50	889.696
6	527	448	4069	53	423	26	224	1051.35	732.024
7	563	476	4107	56	420	31	249	1007.61	693.108
8	571	485	3783	57	473	29	209	688.80	682.014
9	537	457	4076	54	439	26	202	783.80	727.538
10	565	480	3682	57	489	28	197	481.47	450.774

For topic identification, a new $\bar{\mathbf{W}}_{326385\times510}$ 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 for Tamil term frequencies. From $\bar{\mathbf{W}}_{326385\times510}$, $\mathbf{S}^{-1}_{100\times100}$ and $\mathbf{U}_{326385\times100}$ matrices, $\mathbf{V}^{\mathbf{T}}_{100\times510}$ matrix is obtained using eq. (8). Using cosine similarity, topics are identified for the selected 10 documents. After topic identification, probability streams obtained from $\mathbf{U}_{326385\times100}$ matrix are adapted into their respective topic-specific models. Finally, the domain independent model is adapted with topic-specific models.

Results

After each direct adaptation in domain independent language model, perplexity of the language model has been tested with the same test set. This language model has been used in in-house ASR and WER is obtained in both stages. The results are summarized in Table 2.

Ten topic-specific models have been 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 3. It may be seen that reasonable improvements have been obtained after adaptations.

Similarly, perplexity of the domain independent model is calculated before and after adaptation through topic-specific models. This has resulted in significant improvement of domain independent model. This domain independent model has been used in ASR and WER is obtained before and after adaptation. The results are shown in Table 4.

Table 4. Results of domain independent model adapted with topic-specific models

Details	Perplexity	WER in ASR
Before adaptation	121.71	7.94
After adaptation	115.56	7.81

Conclusion and future direction

Resource deficiency in Tamil has been overcome through CLSA and adaptation. Probability sparseness of content words has also been overcome by applying partial morphology. The results show significant improvement in perplexity and WER obtained before and after direct adaptation with baseline model. The probabilities of content words have been further boosted by adaptation of domain independent model with topic-specific models rather than by direct adaptation.

Both the approaches adapt unigram translation probabilities. In the future, similarity in phrasal or chunking structures can be studied. The noun with cases in Tamil is equivalent to the noun with prepositions in English. Verb groups in Tamil are equivalent to verb and its auxiliaries in English. Prefix with noun in Tamil is equivalent to determiner with noun in English. Similar chunks or phrases can be identified. These cases may be treated as single or separate entities in English and Tamil. Without partial morphology, CLSA approach may be applied to study N-gram possibilities in English documents and higher order adaptations may be done with bigram and trigram probabilities in Tamil trigram model. For the topic-specific models, more and more documents related to trained documents can be projected into reduced CLSA space[18]. For totally diverse documents, CLSA space can be broadened to encompass many more topics and words. Similarly, the alignment model can be modified to accommodate similarities between English and Tamil phrases.

1. Jelinek, F., *Statistical Methods for Speech Recognition*, MIT Press, 1997.
2. Saraswathi, S. and Geetha, T. V., Comparison of performance of enhanced morpheme-based language model with different word-based language models for improving the performance of Tamil speech recognition system. *ACM Trans. Asian Lang. Inform. Process.*, 2007, **6**, Article 9.
3. Och, F. J. and Ney, H., Improved statistical alignment models. *Proc. ACL*, 2000, 440–447.
4. Kim, W. and Khudanpur, S., Language model adaptation using cross-lingual information. In *Proceeding of Eurospeech*, Geneva, Switzerland, 2003, pp. 3129–3132.
5. Kim, W. and Khudanpur, S., Cross-lingual latent semantic analysis for language modeling. In *Proceedings of ICASSP*, Montreal, Quebec, Canada, 2004, vol. 1, pp. 257–260.
6. Kim, W. and Khudanpur, S., Cross-lingual latent semantic analysis for language modeling. *IEEE*, 2004, 257–260.
7. Yik-Cheung Tam, Ian Lane and Tanja Schultz, Bilingual-LSA based LM adaptation for spoken language translation. *Proc. ACL*, 2007.
8. Landauer, T, K., Foltz, P. W. and Laham, D., Introduction to latent semantic analysis. *Discourse Processes*, 1998, **25**, 259–284.
9. Dumais, S., Landauer, T. and Littman, M., Automatic cross-language retrieval using latent semantic indexing. In AAAI Spring Symposium on Cross-Language Text and Speech Retrieval, 1996, pp. 18–24.
10. Berry, M., Dumais, S. T. and O'Brien, G. W., Using linear algebra for intelligent information retrieval. *SIAM Rev.*, 1995, **37**, 573–595.
11. Husbands, P., Simon, H. and Ding, C., The use of singular value decomposition for text retrieval. In Proceedings of SIAM Comp. Info. Retrieval Workshop (ed. Berry, M.), 2000.
12. Kim, W., Language model adaptation for automatic speech recognition and statistical machine translation. Ph D thesis, The John Hopkins University, Maryland, 2004.
13. Rajendran, S., Strategies in the formation of compound nouns in Tamil. *Languages of India*, 2004, vol. 4, p. 6.
14. Rajendran, S., Viswanathan, S. and Ramesh Kumar, Computational morphology of Tamil verbal complex. *Language of India*, 2003, vol. 3, p. 4.
15. Coccaro, N. and Jurafsky, D., Towards better integration of semantic predictors in statistical language modeling. *Proc. ICSLP*, 1998, **6**, 2403–2406.
16. Ronald Rosenfeld, Adaptive statistical language modeling: a maximum entropy approach. Ph D thesis, Carnegie Mellon University, Pittsburgh, 1994.
17. Jerome, R. B., Exploiting latent semantic information in statistical language modeling. *Proc. IEEE*, 2000, **88**, 8.
18. Yik-Cheung Tam and Tanja Schultz, Incorporating monolingual corpora into bilingual latent semantic analysis for cross-lingual LM adaptation. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP '09), 2009.

ACKNOWLEDGEMENT. We thank Central Institute of Indian Languages, Mysore, India for providing Tamil text corpus.

Received 17 July 2009; revised accepted 3 February 2010