Phrase based English – Tamil Translation System by Concept Labeling using Translation Memory

R. Harshawardhan
Department of CEN
Amrita School of Engineering
Coimbatore, India

Mridula Sara Augustine
Department of CEN
Amrita School of Engineering
Coimbatore, India

K.P. Soman
Department of CEN
Amrita School of Engineering
Coimbatore, India

ABSTRACT
In this paper, we present a novel framework for phrase based translation system using translation memory by concept labeling. The concepts are labeled on the input text, followed by the conversion of text into phrases. The phrase is searched throughout the translation memory, where the parallel corpus is stored. The translation memory displays all source and target phrases, wherever the input phrase is present in them. Target phrase corresponding to the output source phrase having the same concept as that of input source phrase, is chosen as the best translated phrase. The system is implemented for English to Tamil translation.

General Terms
Natural Language Processing.

Keywords
Phrase Based, Translation Memory, Concept Labeling.

1. INTRODUCTION
The translation system may be of Rule based, Example based or Statistical approach. Apart from this traditional way of translations, there is a unique system that is constantly emerging in the translation field, known as phrase based translation system. This type of phrase based translation engine can be developed using any of the three approaches as discussed above. The resources available for Statistical Machine Translation (SMT) are very less for Indian languages, especially Dravidian. Moreover Indian languages are more complex to model. The rule based systems are very constrained over handcrafted rules and since the language is continuously evolving, the rules had to be adapting accordingly [1]. Here we go for example based approach and Translation Memory (TM) serves our purpose.

TM is not exactly a translation system but it is made to assist the human translators by providing a coarse translation. TM is used to implement natural language processing (NLP) tools for any languages. We are using OmegaT, open source software based on Fuzzy Matching algorithm which performs exact searches [2] and OmegaT is found to be more transparent than NATools [3]. When exact match for the search text is found, then OmegaT gives all text segments of source and target text, wherever input phrase is found in corpus. When no match is present, the system automatically gives some partially related output text segments based on fuzzy match. In order to aid this matching process for more accurate results, we are including concept labeling. Therefore the text fragments are searched and retrieved in TM along with their concept labels.

2. TRANSLATION OF PHRASES
The bilingual lexicons and bilingual sentences contribute for word by word translation and sentence level translation respectively. Such resources for Indian languages are scarce and make the regular translation systems hard to implement. The phrase based translation solves this problem since it stands as an intermediate between the words and sentences. Current state of the art models in machine translation are based on alignments between phrases [4]. The complexity in machine translation (MT) comes into picture, if the language is having many inflection forms, where we require a morph analyzer to understand the inflections of source language and a morph synthesizer to generate those inflections in the target side. In case of translation of phrases, morph analyzer or synthesizer is not necessary, because we are considering only the n-grams, which are already might be inflected. The output will be having the inflection forms of that phrase, existing in the corpus, even when we search for other inflections, which is sometimes a disadvantage.

The advantage of phrase based translation is that we don't have to worry about the word sense disambiguation (WSD). The same source word may represent semantically varying words with respect to each sense. Choosing the appropriate word for a particular sense involves WSD process [5]. But here we are dealing with unigrams, bigrams, trigrams, etc. There is no need for explicit mapping of individual words according to senses. They carry the same sense information implicitly within the phrase and this information is preserved.

A group of words together represent a different meaning from what they mean individually [6]. These multiword expressions create ambiguity when they are getting translated word by word. For example, the word *'hard disk'* represents a storage device and not exactly the *'tougher disk'*. In case of translation of phrases, we are considering the phrase *'hard disk'* as a whole, so the translation is also a complete phrase. The hard part of translation is translating the idioms and phrases. This problem is tackled easily by the phrase based translation system since the system considers them as a regular group of words. The phrase based translated output may look complicated but still we claim that the translation is understandable and it provides the abstract meaning.

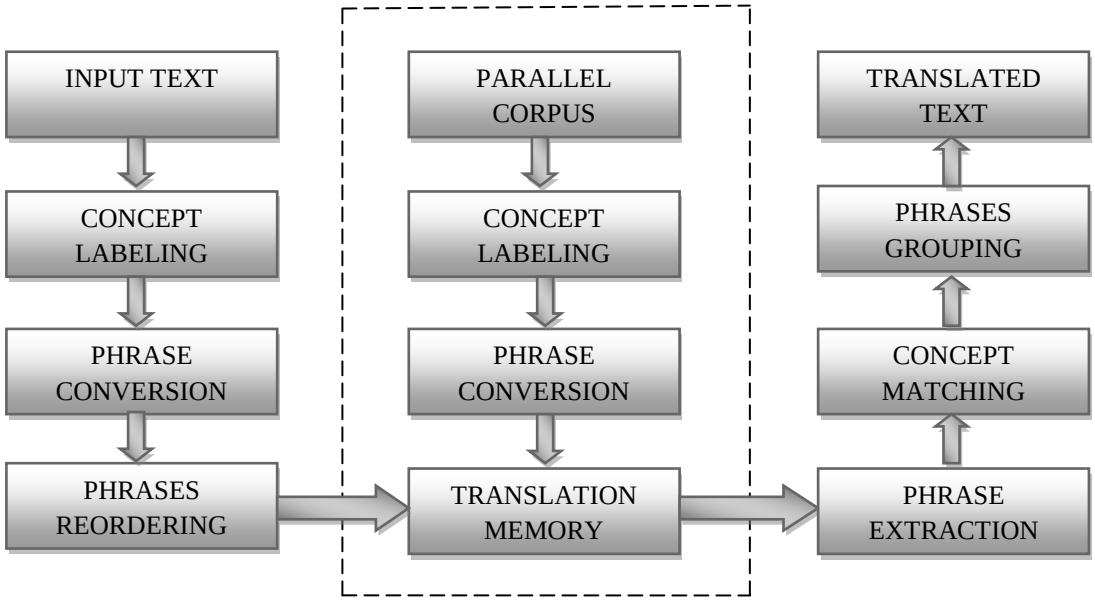

Fig 1: Block diagram of translation system.

3. CONCEPT LABELING

We go for concept labeling because it provides a solution for disambiguation of phrases. The words can be disambiguated according to their domains with the help of WSD system, but phrases are not of this kind. The words in a phrase individually have their own senses. But these words together as a phrase need not possess the same sense. So we have to relate the phrases with some approach. Using N-gram model we can possibly predict the forthcoming phrases, but we cannot relate the phrases. Relating the phrases word by word is also time consuming and not necessary too. We require some information that each phrase could carry, so that the appropriate target phrase could be retrieved exactly in TM. This task is accomplished with the help of concept labeling.

The concept labeling is a recent approach based on the fact that how our human mind makes disambiguation decisions, to interpret the language, especially words. We use both world knowledge and linguistic information for labeling the concepts [7]. Each word in a sentence can be physically realizable with our mind by visual perception. But this knowledge is hidden inside a sentence, which could not be found by a machine until we explicitly represent this kind of information to the machine. This knowledge may include physical entities of the world such as person, object or location, abstract concepts or relationships. For a particular language, even the morphological information, responsible for inflections among the words could be considered for labeling a concept, only if a physical knowledge is present.

4. IMPLEMENTATION IN TM

The block diagram of the complete translation system is shown in Figure 1. The central block within the dotted lines represent the preparation phases of TM. This phase make the system to get prepared for translation and makes it ready for working. This phase is essential because the accuracy of the system relies on the perfection of this part. The outer blocks on both the sides of the middle block, represent the working phase of translation system. Each block of the diagram is explained in detail as follows.

4.1 Preparation Phase

The collection of parallel corpora forms the initial part of preparation phase. We implemented the translation system for English – Tamil language pair and collecting such a bilingual pair is tedious. The information available in web is not sufficient. So we collected the bilingual books of free copyrights like short stories, biographies and encyclopedias containing the parallel texts and we manually typed and aligned the text. The Government of Tamil Nadu academic textbooks of schools provide good parallel corpora which is also typed and aligned manually.

For better understanding in TM, the English – Tamil dictionary is loaded along with their POS categories. The idioms and phrases of English and Tamil, and proverbs of English and Tamil are also included in the corpora to improve the system accuracy since their translations vary from usual translations. The alignment of parallel corpora should be proper since one or more source sentences may be mapped to one or more target sentences [8]. The terminology database containing most of the bilingual technical terms in domain wise, is also included in our corpora. On considering the advantages of a phrase based translation systems, Google released the n-grams data from Google books [9]. The corpus is downloadable for free and has millions of n-grams, including unigrams, bigrams, trigrams, tetragrams and pentagrams. We manually translated a part of these n-grams and we prepared our own bilingual n-grams and included in the corpora.

Table 1. Working of different TM systems for a sample input

TM System	Concept Labeling	Sample Input	Probable Fuzzy Matched Output
Regular	No	I had Salmon in the bank	I had pen in the pocket
Regular	Yes	I had Salmon in the bank – (Concept: Eat)	I had bread in the morning – (Concept: Eat)
Phrase based	No	I/ had Salmon/ in the bank/	I/ had money/ in the bank/
Phrase based	Yes	I – (Concept: 1st person)/ had – (Concept: Eat)/ Salmon – (Concept: Fish)/ in the bank – (Concept: Land; Water)/	I – (Concept: 1st person)/ had – (Concept: Eat)/ Mackerel – (Concept: Fish)/ in the beach – (Concept: Land; Water)/

The next part of preparation phase is concept labeling the corpus followed by breaking the sentences into phrases. Sometimes the source sentence itself may carry the concept information to be labeled. For example, the prepositional phrases provide the information about place, time, etc. The concept information is extracted using a Stanford parser by finding the prepositional phrase and the concept is labeled with the sentence. This process of labeling is carried out automatically for all sentences of parallel corpora in source side. For sentences that do not implicitly carry the concept information, we explicitly labeled the concepts on source side alone, based on our visual perception.

We are not labeling the concepts for target side, because in TM, the target text accompanies with the labeled source text and there is no need for labeling them. During parsing, we convert all the source sentences into phrases. The concept labels are passed from sentences to phrases of that particular sentence and now the phrases are included in the parallel corpora. When we are translating a phrase using TM, the output we require should also be a phrase, which is made possible by converting the target sentences of corpus converted into phrases.

The last stage of preparation phase involves the loading of parallel corpora in TM system. For OmegaT, the corpora should be converted into *'tmx'* format [10] and the process is automated. So the entire corpora (sentences of textbooks, idioms and phrases, proverbs and the converted phrases from sentences) is converted into *'tmx'* format and loaded into OmegaT along with their concept labels. English-Tamil dictionary is developed on *'StarDict'* platform and loaded into OmegaT separately.

4.2 Working Phase

We see how the system works for a sample input. If the system is given a paragraph as input, a generalized concept for the whole paragraph like 'location' is found and that concept is labeled to the whole paragraph. Then the paragraph is separated into individual sentences and the concept is found for each sentence like 'time' information and these sentence level concepts are labeled to all the sentences of paragraph along with the concepts of paragraph. Each sentence is broken into phrases using Stanford parser and each phrase carries the concept labels of sentences as well as paragraphs. These phrases are fed into the TM system in a reordered fashion with simple rules of target language. Here we should note that the reordering among the phrases is done instead of reordering among the words.

Reordering is included in order to improve the quality of translation and it is optional. OmegaT searches through the corpus and if input phrase is found, it gives all source text fragments containing the input phrase along with associated target text, dictionary terms, POS category and domain for every word in a given phrase. We search and find the concept label that exists in common between input and output source phrases. We retrieve the output target phrase corresponding to output source phrase that has same concepts of input source sentence and source paragraph. This retrieved phrase is taken a translated phrase.

The process is repeated for all the phrases of given sentence and put all output phrases together in the target side with respect to the reordered source sentence. All these translated target sentences are combined in a regular order to form the translated paragraph. Reordering of sentences is not needed when they are combined as a paragraph.

Suppose if the input phrase is not found in the corpus, the fuzzy matching algorithm of OmegaT searches for similar phrases in entire corpus and display all similar phrases. If no such similar phrase exists in corpus, then the algorithm breaks the input phrase into smaller n-grams and searches in corpus again. The target phrase of the fuzzy output can be considered as an approximate translation, if the fuzzy match possesses same concept labels as that of input text. Then we retrieve and combine all target outputs in a similar way mentioned before.

We don't require full sentences for translation of phrases. But we included such sentences in the corpora, to enhance fuzzy matching. If the input is given as a single sentence, the paragraph splitting part could be avoided. Similarly if the input is given as phrases directly, sentence breaking with the help of parser is also not required. The concept for the phrases if exist, will be labeled automatically according to the phrase, else, the concept has to be labeled manually.

5. COMPARISON OF TM SYSTEMS

Table 1 compares the working of four kinds of TM systems: Regular TM systems with and without concept labeling and phrase based TM systems with and without concept labeling.

Consider the sample input *"I had Salmon in the bank"*, which states that *"I ate fish in the river bank"*. Let us assume that our corpus does not contain exactly the same sentence and let us see the probable fuzzy matched outputs for all four kinds of TM.

The first system is a regular TM without concept labeling and the output would probably be *"I had pen in the pocket"*. The system finds a fuzzy output that has a same structure as input but deviates much from the input in meaning.

The second system is a regular TM with concept labels. For the given sample input, the concept is chosen as *"Eat"* and the corresponding output might be a sentence that has something to do with eating. The probable output could be *"I had bread in the morning"*. The output sentence might be similar to the input in *'Eating'* sense, but it is not a good match for translation in TM.

The third system is a phrase based one without concept labeling. Fortunately for a given sample input, the phrase *"in the bank"* is already exists in the corpus. The output would then be *"I had money in the bank"*, which is a better output than the first one. But the context it refers the bank belongs to a *"savings bank"* category, which is also not acceptable as a good match.

The fourth and the final system we consider is the phrase based TM with the concepts being labeled for each phrase with the help of parser. The possible output would be *"I had Mackerel in the beach"*, where each phrase of output shares the same concept as that of input phrases, as given in Table 1. The output is not exactly the same match for translation we expected. But still the output is considered as an acceptable fuzzy match and the translation of which, will be a much better one than the other TM systems we considered earlier, because the output sentence preserves the meaning of the input sentence.

It is known that for all four TM systems, the fuzzy outputs do not limit with the probable outputs we considered earlier and always there will be numerous matches. We consider the worst possible case of fuzzy matches for our convenience to get a better understanding of all the systems.

6. RESULTS

The accuracy of translation depends on the size of parallel corpora. The greater the corpus size the lesser the fuzzy outputs are. The output of OmegaT carries a lot of information including the dictionary, domain, POS category, proverbs, idioms and phrases. These features are included in order to provide additional information about translations to aid human translators using our system. We developed the system with 50,000 English – Tamil parallel sentences, 5000 proverbs, and 1000 idioms and phrases, with a dictionary containing more than 2,00,000 technical words and 100,000 general words. We are able to get the accuracy of 70% while testing the system with standard phrases.

The graph in Figure 2.a evaluates the performance of regular and phrase based TM systems by measuring the percentage of translation accuracy with respect to a given corpus size. The graph shows that the accuracy in translation with TM increases with the size of the corpus in case of both the TM systems. Regular TM shows a steady rise in the curve whereas phrase based TM graph rises in a higher rate than that of regular TM. The increase in translation accuracy is because of the fact that, when the number of sentences in corpus increases, then correspondingly the number of phrases also increases at a higher rate. With more number of phrases, the possibility to get a good match for translating the phrases is more than with the small corpus. Also for regular TM with huge corpus, the good translated matches are better than those with the small corpus. But still the accuracy of translation is less compared to phrase based TM because the sentences in regular TM are not increasing in the same rate of phrase based TM. The phrases increase at a much higher rate than that of sentences because each sentence carries more than a single phrase. Therefore for increased corpus size, more number of phrases will be available for matching than the sentences of regular TM. From this graph one can deduce that the phrase based approach works well with TM.

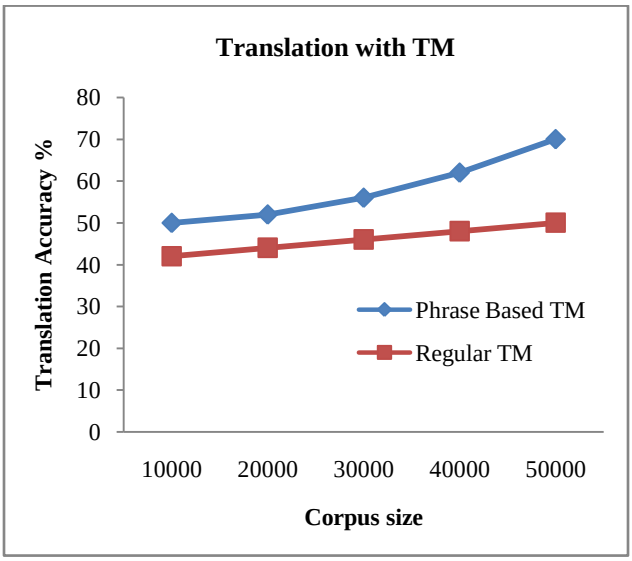

Fig 2.a: Graph showing the performance of Regular & Phrase based TM.

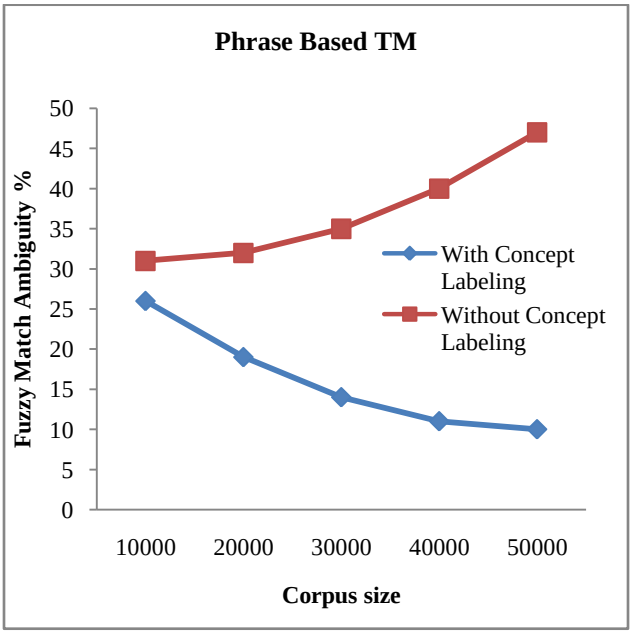

Fig 2.b: Graph showing the ambiguity of Fuzzy Matches for phrase based TM

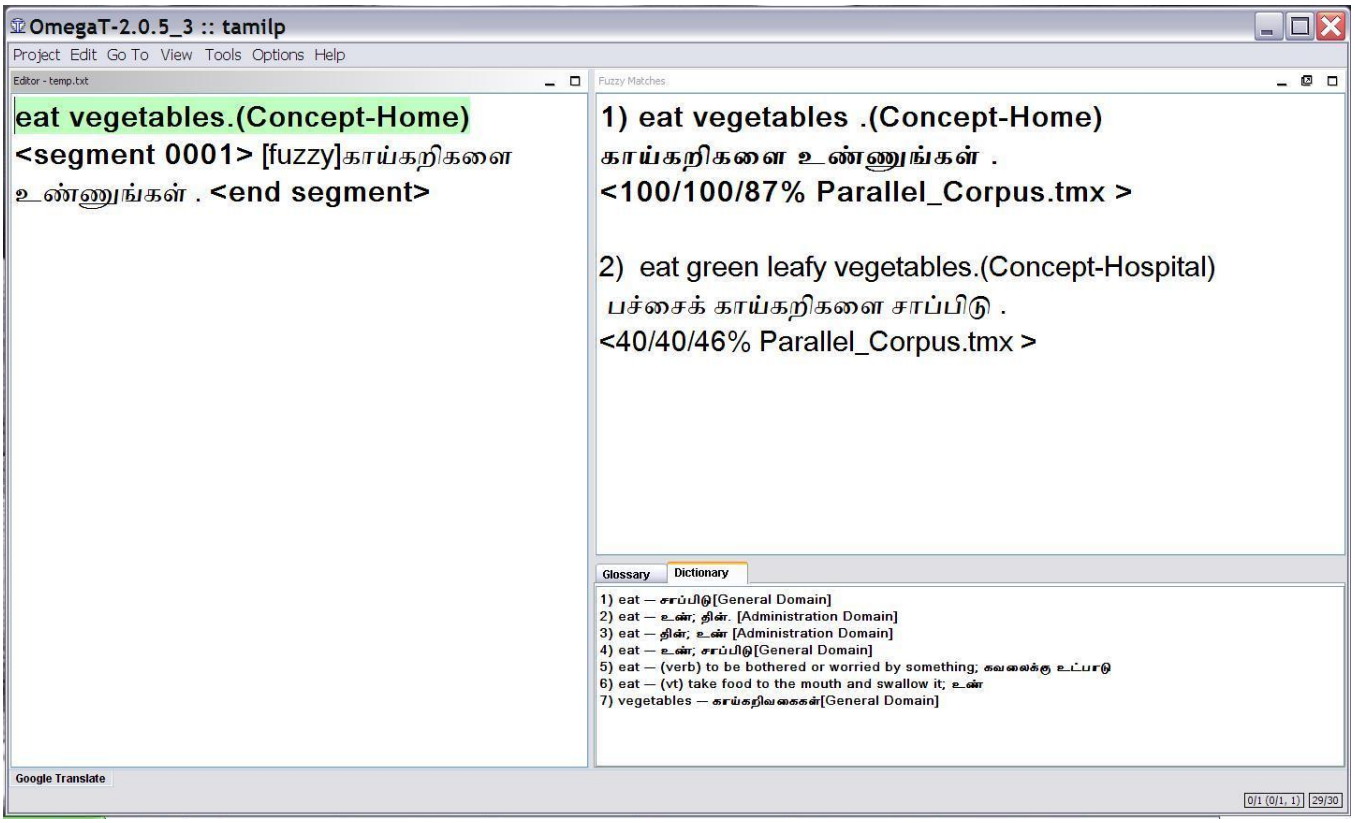
OmegaT-2.0.5_3 :: tamilp
Project Edit Go To View Tools Options Help
Editor - temp.txt
eat vegetables.(Concept-Home)
<segment 0001> [fuzzy]காய்கறிகளை
உண்ணுங்கள் . <end segment>
Fuzzy Matches
1) eat vegetables .(Concept-Home)
காய்கறிகளை உண்ணுங்கள் .
<100/100/87% Parallel_Corpus.tmx >
2) eat green leafy vegetables.(Concept-Hospital)
பச்சைக் காய்கறிகளை சாப்பிடு .
<40/40/46% Parallel_Corpus.tmx >
Glossary
Dictionary
1) eat — சாப்பிடு[General Domain]
2) eat — உண்; தின். [Administration Domain]
3) eat — தின்; உண் [Administration Domain]
4) eat — உண்; சாப்பிடு[General Domain]
5) eat — (verb) to be bothered or worried by something; கவலைக்கு உட்பாடு
6) eat — (vt) take food to the mouth and swallow it; உண்
7) vegetables — காய்கறிவகைகள்[General Domain]
Google Translate
0/1 (0/1, 1)
29/30

Fig 3: Screenshot showing the sample output of OmegaT.

The graph in Figure 2.b shows the significance of concept labeling for phrase based TM systems. Previously we discussed regular and phrase based TM systems for an exact match for translation. Now we consider the phrase based TM system alone with the assumption of not getting the exact match, the output will now be the fuzzy matches. From the graph in Figure 2.a, it is known that the number of phrases increase with the corpus size. Therefore the occurrences of fuzzy matches will also be more and choosing the best fuzzy match would be an issue. In phrase based TM without any concept labeling, the ambiguity of fuzzy matches increases with the corpus size. Numerous fuzzy matched phrases would there in TM which makes the selection of the best fuzzy matched phrase for translation becoming more ambiguous. This problem is easily solved by concept labeling with the implicit saying *"The more the corpus size the more delicate the concepts are"*. The concepts will get refined with the increase in corpus size. For example, with small corpus, a particular phrase may be labeled with the concept *'home'*, but when the corpus increases, the same concept may not be a good choice, because the same concept might be carried by more number of similar phrases. Therefore the concept has to be more clear and particular about that phrase. So the concept *'home'* should be replaced by either *'drawing room'* or *'kitchen'*, etc. The number of concepts also tends to increase with the corpus size and the fuzzy matches carry more specific concepts which are easy for alignment. The graph in Figure 2.b clearly shows that for a phrase based TM with concept labels, the ambiguity in choosing the best fuzzy outputs decreases with the increase in corpus size, due to increase in the number of concepts. This makes the concept labeling much suitable for phrase based TM systems.

The screenshot of a sample output is shown in Figure 3; the input phrase is *'eat vegetables'*. The concept *'home'* is found based on the input sentence *'I eat vegetables for lunch'*. Note that the phrase *'eat vegetables'* is present in both the outputs of TM. The best matching phrase is found based on concept that is common between the input source phrase and the output source phrase. That is, the *'Concept-Home'* is common for input and output phrases.

7. CONCLUSION

The translation memory is a simple tool that stores and retrieves data. Anyone can implement this system for any language, as it is irrespective of the language. But when utilized in a proper manner, this simple tool would serve as an indispensable material for machine translation. The system which we developed can be used as a model for SMT. The drawback of the system, is that the most of the sentences are from government school texts and short stories for children, the system is constrained over a limited domain. The only way to improve the accuracy is that to increase the resources of parallel data for Indian languages, especially Dravidian in web, as an open source. We are trying to employ this system as a free web

application. Millions of n-grams are uploaded in our site for people to contribute for the translation of n-grams and can be downloaded and utilized for research purposes. The completed n-grams will be added to the system and stored in TM. Whoever wanted to contribute for translating the n-gram corpus, can refer this TM. TM could give them translation of n-grams if present in any sentence of the corpus. The additional information of world knowledge from concept labels, dictionary, domain wise technical information and POS category of the words together give an abstract idea about the n-gram being searched in TM.

8. APPLICATIONS

The TM itself serves as an aid for human translators by providing the abstract output. The system is robust enough to handle huge paragraphs and large sentences that are required for high end applications of NLP. The translation system implemented using TM is scalable and can be extended for many number of applications. The phrase based text system can be further developed into a phrase based speech system to be used for differently-abled. The phrase based translation system can be employed in mobile phones or any hand held devices to assist tourists. The system can be used for pedagogical purpose to teach grammar among school children.

9. FUTURE WORK

The future work includes:

- The same system can be implemented with massive database.
- The system can be extended for multilingual translations.
- Language processing tools can be included to improve the translation.
- The concept labeling can be automated through machine learning approach.
- TM can be modified and used for word alignment in SMT.
- Semantic and syntactic information of phrases can be included.

10. REFERENCES

[1] Alon Halevy, Peter Norvig, and Fernando Pereira. 2009. The Unreasonable Effectiveness of Data. Article. In IEEE Intelligent Systems, (vol. 24 no. 2).

[2] Dmitri Popov. 2005. Translating With OmegaT. Internet Article. Available: http://www.linux.com/articles/42532.

[3] Alberto Manuel, Brandao Simoes. 2004. Parallel corpora word alignment and applications. Master's thesis. Department of Computer Science, University of Minho, Braga, Portugal.

[4] Philipp Koehn, Franz Josef Och, and Daniel Marcu. 2003. Statistical phrase-based translation. In Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics (HLT-NAACL), pages 127–133.

[5] Eneko Agirre and Aitor Soroa. 2009. Personalizing PageRank for Word Sense Disambiguation. In Proceedings of the 12th conference of the European chapter of the Association for Computational Linguistics (EACL), Athens, Greece.

[6] Whitney St.Charles. 2008. Noun Phrase Extraction – An evaluation and description of current techniques. Departmental Honors Thesis, Department of Computer Science, The University of Tennessee, Chattanooga.

[7] Antoine Bordes, Nicolas Usunier, Ronan Collobert, Jason Weston. 2010. Towards Understanding Situated Natural Language. In Proceedings of the 13th International Conference on Artificial Intelligence and Statistics (AISTATS), Chia Laguna Resort, Sardinia, Italy. Volume 9 of JMLR: W&CP 9.

[8] Philipp Koehnn. 2010. Statistical Machine Translation. Cambridge University Press.

[9] Jean-Baptiste Michel*, Yuan Kui Shen, Aviva Presser Aiden, Adrian Veres, Matthew K. Gray, William Brockman, The Google Books Team, Joseph P. Pickett, Dale Hoiberg, Dan Clancy, Peter Norvig, Jon Orwant, Steven Pinker, Martin A. Nowak, and Erez Lieberman Aiden*. Quantitative Analysis of Culture Using Millions of Digitized Books. Science (Published online ahead of print: 12/16/2010).

[10] Susan Welsh and Marc Prior. 2009. Omega T for cat beginners. Official documentation of Omega T.