Towards a Pictorially Grounded Language for Machine-Aided Translation

Samir Kr. Borgohain[1], Shivashankar B. Nair[2]
[1] National Institute of Technology Silchar, Department of Computer Science & Engg., Assam, India 788010
[2] Indian Institute of Technology Guwahati, Department of Computer Science & Engg., Assam, India 781039
samirborg@gmail.com, sbnair@iitg.ernet.in

Abstract

In this paper, we present a translation system that can aid people not well versed in each other's languages to communicate their ideas through an intermediate pictorial knowledge representation system called the Pictorially Grounded Language (PGL). Machine-aided translation augmented by PGL culminates in symbols of both the languages – the source and the target - being grounded on a common set of images and animations. The system discards the conventional intermediate language representation and substitutes it with the PGL. Since PGL is a graphic language comprising of a sequence of images and animations and since both the parties can see these being rendered, they remain convinced that the ideas emanating in one language (source) are the same as that being generated in the other (target), thus ensuring reliability of translation. The mechanism of translation is also scalable and can be used to translate a language into a larger set of languages while at the same time is capable of preserving their inherent meanings. The paper also presents the results of translations augmented by the corresponding PGL versions that tend to improve the comprehension in the target language.

Keywords

Natural language; symbol grounding; machine-aided translation; graphic language.

1 Introduction

Language is a combination of sounds, words and grammar that aid in conveying the meanings, thoughts or ideas of an intended user. There are innumerable languages in the world today and a varying number of native speakers who make use of them. The communication using natural languages becomes difficult between speakers when each of them knows a language which the other does not. Translation, under such circumstances has to be effected using a third party who is conversant in both these languages. Finding or generating such a third party can pose to be a big problem if these languages are not popular. Many linguists are skeptical about translation and consider that it to be nearly impossible because each language is the subjective reality of its native speakers (Hutchins 1986). Computer systems have been also used to aid in machine translation. However the problem

persists as a computing system cannot comprehend the semantics of a language. The Chinese room experiment carried out by Searle (Searle 1980) demonstrates the way in which computer systems process a language. As per his theory, linguistic words are merely meaningless symbols. Harnad (Harnard 1990) emphasized the need for symbol grounding. Even statistical machine translation systems (Koehn 2005), which among others exploit parallel corpora to learn mappings between two different languages, fail to address the problem of symbol grounding. Quine (Vogt et. al 2007) describes a situation when a person needs to communicate with a tribal whose language is not known to us. Under such conditions we are forced to find means of making sense out of the utterances and gestures that help us ground his/her language. Quine claims that in such a situation, it is impossible, to be absolutely certain of the meaning of the utterances made by that tribesman. For example, if such a tribesman sees a rabbit and says *gavagai*, ambiguity can creep in due to the following - Is he referring to rabbit as a whole, or to a specific part of the rabbit, or to some temporal aspect related to the rabbit? If we consider the symbol grounding problem, there could be practically an infinite number of conceptualizations for such a situation. Quine also mentions that one can form manuals for translation. The observer examines the utterances as parts of the overall linguistic behavior of the individual and then uses these observations to interpret the meaning of all the other utterances.

No machine translation system is without side effects. While a human translator may be trusted, his/her software equivalent cannot be. One can never be sure of the quality and reliability of such translations. Assuming the reader understands only English, a translation system could be tested for its reliability by performing translation of an English sentence to another language (e.g. Spanish) and then retranslating the same back to English. In Table 1, we cite the case of two popular translation systems available on the web and use them to perform such a translation to bring out their effectiveness. Though the meaning seems to be conveyed in a way, in the translations, it may be noted that the reverse translation gives odd results in some cases making the English user unsure of whether the correct semantics were actually conveyed to the Spanish reader. *Swinging* using a rope and *moving* with a rope may not convey the right meaning just as *picking* and *choosing* a pen. Symbolic translations of such actions augmented with images, or better still, animations, could act as a means for their correct and unambiguous interpretation. The table also reflects, to some extent, on the quality of the Spanish to English translator. However, the main point being emphasized here is that there is no way either of the speakers (or readers) can say for sure whether the meaning of the sentence is correctly and unambiguously conveyed in the other language. Ensuring this is an important aspect of communication. The problem becomes adverse if we need to convey to a person in a language for which a translator is not available.

English	Spanish	Back to English	English translation of the Spanish sentence	Spanish	Back to English	English translation of the Spanish sentence
Language is used for the purpose of communi-cation among human beings and some animals	Translator: Babylon (http://translation.babylon.com)		Translator: Human	Translator: Babelfish (http://babelfish.yahoo.com)		Translator: Human
	El lenguaje es utilizada para el fin de la comunic-acion entre los seres humanos y algunos de los animals	The language is used to end the communi-cation between humans and some of the animals	The language is used for the purpose of communic-ation between the human beings and some of the animals	La lengua se utiliza con el fin de la comunic-ación entre seres humanos y algunos animals	The language is used with the purpose of it communi-cation between human beings and some animal	The language is used with the purpose of communica-tion between human beings and some animals
The boy is swinging using a rope	Boy se mueve con cuerda	Boy moves with rope	Boy moves with rope	El muchacho está haciendo pivotar con la cuerda	The boy is making pivot with the cord	The boy is swinging with the rope
Give me an apple	Me da una manzana	I got an apple	Give me an apple	Déme una manzana	Déme an apple	Give me an apple
Pick a pen	Elige un boligrafo	Choose a brown	Choose a pen	Escoja una pluma	It chooses a pen	Choose a pen
The boy is walking	El muchacho esta caminando	The boy is walking	The boy is walking	El muchacho está caminando	The boy is walking	The boy is walking

Table 1. Translation of some sentences from English to Spanish and their re-translations using two popular online translation systems together with their actual meanings.

2 Survey on Machine Translation Systems

Machine translation (MT) is the process of translating the contents from a source language to a target language with the help of computers. The concepts or thought processes of the source language should be preserved while converting it to the target language. MT systems may be categorized based on whether human intervention is required before, during or after the

translation process. These systems are also broadly categorized as - human-assisted translation and fully automatic translation. A number of strategies (Hutchins 1986) have been employed to achieve machine translation. They are chosen and applied based on the syntactic and semantic nature of both the source and target languages. MT strategies fall into three broad groups (Chellamuthu 2002) 1) Direct Translation strategy, 2) Transfer strategy and 3) Interlingua strategy. In the direct translation strategy, the source language text is analyzed and directly translated to its target language equivalent through a series of operations. It neither uses an intermediate language nor a parsing system. The output of this system depends on a codified dictionary, pre-specified sentence patterns and on morphological analysis. The Georgetown machine translation system (Slocum 1985) uses the direct strategy. The primary objective of this system was to translate Russian text to English. The strategy employed by the Georgetown MT system was simple word-for-word replacement, followed by a limited amount of transposition of words to result in something vaguely resembling English. The design of the system was monolithic in nature. The Systran machine translation system (Slocum 1985), an improvised and less complex version of the Georgetown system, is characterized by its use of a hybrid 'direct-transfer' system wherein the programs for structural analysis and synthesis are largely independent. The main translation processes are driven by the source to target language dictionaries, as in direct systems (Hutchins 1986), (Hutchins 1995). In the case of the transfer method the source language text is analyzed and transferred into an intermediate language called a meta-language with the help of the target language lexicon and then reconstructed before transforming the sentences in accordance with the syntax of target. The translation system (Chellamuthu 2002) due to the research group called GETA in Grenoble University, France, is based on such a method. The interlingua strategy uses an intermediate or universal language to effect the translation. The method uses a knowledge base consisting of rules pertaining to the language. This method employs a universal language which is independent of the natural languages being translated and consists of several stages that include analysis of the text for conceptual representation, providing contextual world knowledge through an inference mechanism and reproduction of the language-free representation of the source sentences into target language sentences. The interlingua based machine translation approach (Hutchins 1986), produces crude 'pidgin' translations. Ceccato (Hutchins 1986) describes the development of an interlingua based conceptual analysis of words and their possible correlations with other words in texts.

The Statistical machine translation (SMT) (Lopez 2008) is another choice for machine translation. The SMT applies a set of rules to transform source language text to its corresponding target language text. Generally, the set of rules are extracted automatically from a parallel corpus. A fast translation system can be built with this technique for a new language pair without having deep knowledge about the new language. The choice of a model is the prime task for carrying out translation using SMT techniques. The word-based model (Brown et al. 1990) developed by IBM uses a word as the basic unit. An enhancement to the word-based model is the phrase-based model. The model chooses a contiguous sequences of words commonly termed as a *phrase* as the basic unit for translation. The translation involves a sequence of steps that consists of splitting a given sentence into phrases, translating the phrases and finally performing a re-ordering of the translated phrases. The above two models are described by using a formalism called the Finite State Transducer (FST). Synchronous context-free grammars (SCFG) (Lewis et al. 1968) have also been a choice for statistical machine translation. This CFG closely resembles the linguistic syntax of a language. CFG based models include *bracketing grammar* (Wu 1996) and *hierarchical phrase-based translation* (Chiang 2007). The former model exploits the linguistic syntax of a language while the latter combines the insights of the phrase-based model with the syntactic

structure. Although FST and SCFG based models are the popular choice for SMT, there are other models too that cater to SMT. Syntactic Phrase based model (Koehn et al. 2003) is one such model that uses tree transducers, which describe operations on, tree fragments rather than strings. An alternative linguistic model for SMT is also proposed where SCFG can be constructed using dependency grammar (Alshawi et al. 2000).With around 24 official languages in India, the machine translation scenario for Indian languages is challenging (Bharati et al. 1999, tdil.mit.gov.in/newIndexSept01.htm). Research groups from across this country have attempted to build machine translations systems for various Indian languages. A research group from Tamil University, Thanjavur, has built a machine translation system using the direct approach to translate Russian documents to Tamil. The National Centre for Software Technology has come up with a web based service for English to Hindi translation. A group at the Indian Institute of Technology Kanpur has developed *Anglabharati*, a multi-lingual translator between English and Indian languages. *Anusaraka*, which is yet another attempt at machine translation, uses Paninian grammar to aid in the conversion from one Indian language to another.

Though a considerable amount of work has been attempted on machine translation (Hedden 1992) a robust system is still not in place. Some of the main reasons behind this have been summarized below:

- Cost: Building a successful machine translation system involves experienced manpower in terms of both linguists and computer professionals, and related hardware and software. There are other hidden costs of making relevant dictionaries for each of these languages.
- Quality: Such systems have been known to produce humorous translations which indirectly reflect on the quality of translation. Though one may argue that human translators are also prone to committing such mistakes, fixing problems of this kind in the natural world is far easier, simpler and faster than in the computational world.
- Poorly-formed source language text: If the source language text is poorly written, then the system will face difficulty in translating it correctly while in case of a human translator the same can be comprehended to a better extent and a fairly accurate translation can be still be achieved.
- Availability of input in a standard form: In many real-world scenarios, the source language may not be available in the form of text. Instead the same may be in the form of images. Although OCRs can solve this problem to some extent, the addition of such modules contributes to raise the cost of translation.

The direct approach is simple and allows one to exploit knowledge about the particular language pair being dealt with, while developing a translation system, but has the disadvantage that we need to create a new system for every new language pair. Thus if we have *N* languages, *N(N-1)* translators need to be created. The interlingua based system could be considered as the theoretically best, since it requires only *N* analyzers and *N* generators. However it requires the creation of a very generic interlingua, which is obviously a difficult task. A good understanding of all the concerned languages needs to be assimilated. The transfer approach lies between these extremes, and is the most widely used one in practice. The statistical approach can create a translator for new language pairs with adequate training data quite easily, but it does not deal explicitly with syntax. For effective machine translation, the prime requirement is a language that is universally understood. If such a language exists or could be made to exist then it could well serve as an interlingua and increase the reliability of machine-aided translation systems. In this paper, we present the use of a more versatile and universally comprehensible pictorially grounded interlingua that can cater to the larger set of languages and thus aid in machine translation.

3 Pictorially Grounded Language

The basic objective behind this work was to build a translation system that can aid people not well versed in each other's languages to communicate their ideas effectively through an intermediate pictorial knowledge representation system called the Pictorially Grounded Language (PGL). Translation augmented by PGL culminates in symbols of both the languages – the source and the target - being grounded (Harnad 1990) on a common set of images and animations. The system discards the conventional intermediate language representation and substitutes it with the PGL. Since PGL is a graphic language comprising of a sequence of images and animations and since both the parties can see these being rendered, they remain convinced that the ideas emanating from one language (source) are the same as those being generated in the other (target), thus ensuring reliability of translation.
Pictorially Grounded Language based translation mechanism can ease the transfer of thought processes from a source to a target language. Translation using PGL is based on an image library which enables the linguistic symbols of a language to be grounded onto images and animations. At present we are not directly concerned about the grammatical issues of the target language but concentrate on how we can reliably convey the meanings intended in the source language to a person comprehending the target language. When we assimilate linguistic symbols in an orderly manner, it is possible to express our ideas. However these linguistic symbols need to be grounded so that they portray their inherent unambiguous meanings when presented in the target language.

Unlike in the development of the conventional interlingua used for translation amongst different languages, one need not study the set of languages to generate the PGL. Further, all users can comprehend PGL since the images and animations rendered, unlike symbols used in conventional methods, allow for intrinsic grounding of the meanings within a user's brain. PGL can also facilitate word sense disambiguation arising at the initial stages of translation. For instance in the query –
"*Where is the bat?"*
the set of images that may appear alongside the word *bat* are –

The user can now choose between one of the two meanings for the word *bat* (viz. the mammal or the wooden club) and ensure that the correct meaning is reflected in the target language. When the query is translated into some other language wherein the equivalent for the word *bat* has also multiple meanings, the target language user remains unconfused as the corresponding image guides him to ground his/her concepts onto the correct and intended meaning of the word.
In the following sections, we show how the grounding of nouns and verbs is realized in the PGL based translation strategy.

3.1 Grounding Symbols in PGL

We have currently used PGL to ground the linguistic symbols that represent the nouns and verbs in a language. Cangelosi (Cangelosi 1999) too showed a significant tendency to evolve

compositional languages made up of verb-noun messages. The language proposed by them consists of two verbs and two nouns. Nouns are basically linguistic entities that are related to visual inputs while verbs are those that are related to actions. In PGL we use them as described below –

(i) Nouns**:** The linguistic symbols that represent the nouns in the input text are grounded by associating them with their corresponding images that symbolize their equivalents in the real world. Such an association is not limited to merely tagging the images with symbols. Some related attributes or properties within the image have also been tagged with symbols. A typical example of a noun being grounded along with its attributes, using an image is shown in Figure 1 and Figure 2. The former depicts some of the nouns that have been labeled. Each of the image objects is marked with ovals and their labels are provided alongside. The attributes (polygon, name, objectID, etc.) of the image object are enclosed in a bounding box within the XML code and shown in Figure 2.

(ii) Verbs: The linguistic symbols that represent verbs in PGL are grounded by a set of actions through animated GIF files. We envisage to represent and ground symbols in a language to their respective image objects that satisfy the properties of that image.

As mentioned, our idea of translation is to convey the meaning of a concept in one language, say L1, to a person who comprehends a language L2, where L1 and L2 are mutually exclusive in terms of the symbols they comprise. Such translation cannot be reliably and unambiguously effected if we use conventional symbol based machine translation methods for reasons cited earlier. Thus rather than present the concept in L1 in a mere textual or symbolic form in language L2, we also augment it with a sequence of images and animations alongside to aid the person to correlate and make more sense by using the information they provide. The image sequence can be seen both by the person who conveys the concept in L1 and by the person who comprehends the same in L2. This allows both parties to visually verify whether the translation of the concept is visually correct thereby increasing the reliability and correctness of the transformation into the other language. Grounding, of course remains intrinsic within the communicating parties and is not achieved by merely two different sets of symbol sequences, one in each language. Instead it is affected by the same sequence of images and animations that universally convey the same meaning to both the parties. Thus the final grounding is achieved by the image sequences and animations generated and rendered by the PGL system.

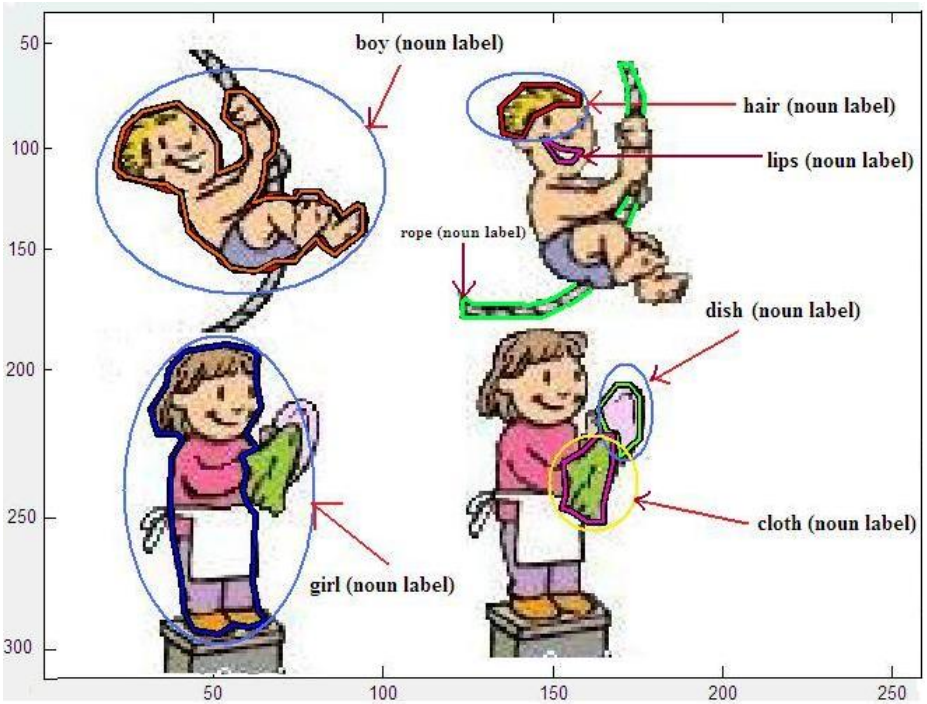


Figure 1. Noun labels along with their representative objects (Figures along the X and Y axes indicate the width and height in pixels.)

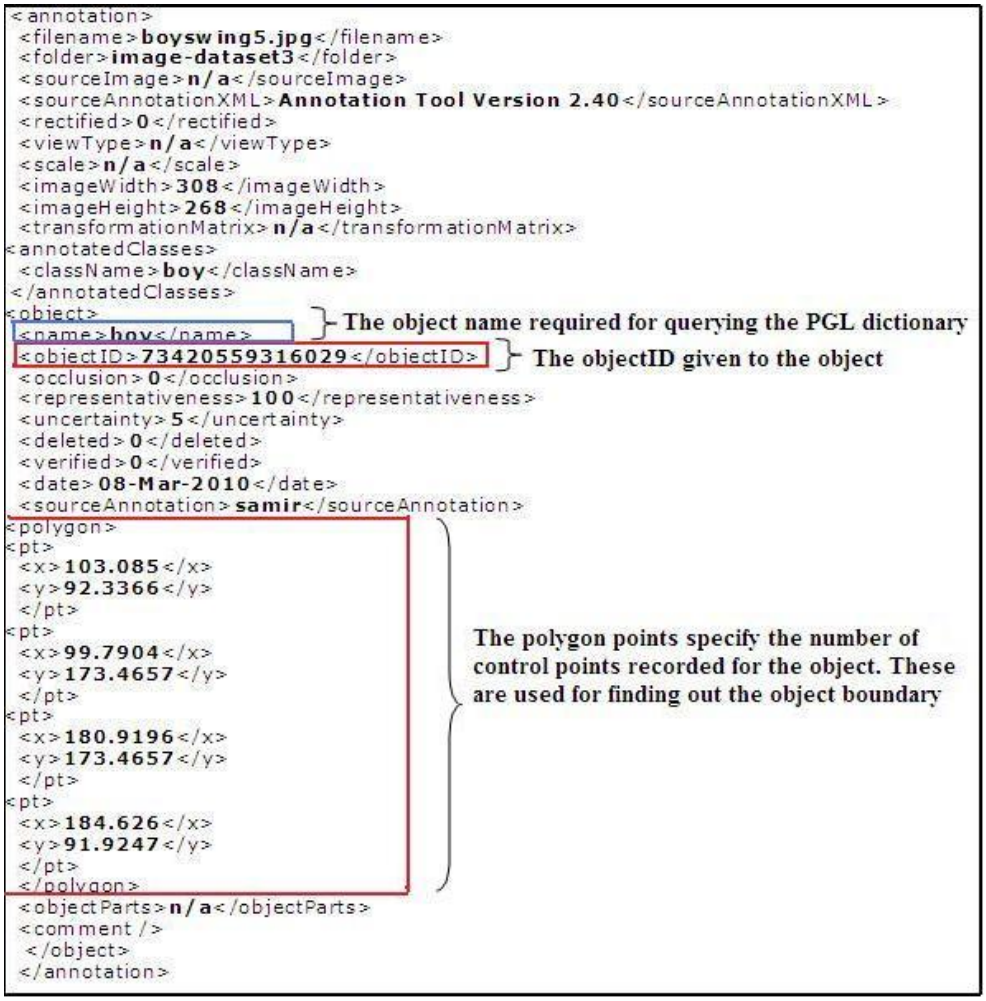


Figure 2. The attributes of the image object are shown within the XML code. The descriptions of a few attributes of the image object are also given.

3.2 Advantages of PGL

The use and potential benefits of a PGL augmented translation system are enormous. A few scenarios, apart from conventional machine translation, where such a PGL augmented translation can work as a novel aid have been enumerated below.

(i) As a language translation verifier:

Since the intermediate PGL is known and seen by both parties involved in the communication, they can verify whether the conveyed meaning tallies with the intrinsically grounded meanings within their respective brains. This makes the translation more reliable.

(ii) As an alien language teacher:

Since concepts like nouns and verbs are pictorially grounded using relevant images and animations, the learner could translate from a known language and read the same in the other unknown language. Since the use of PGL takes care of word sense disambiguation at the initial level the user is assured that the translation does not lead to other meanings in the target language. Realizing such a system may require some amount of speech synthesis for the alien language. Rudimentary learning of the new language could of course be realized almost without any aid from a tutor. Such a system would be beneficial for globe trotters who need to quickly pick up and convey in the local language or dialect without much stress on the syntax. Sentences from stories input to the translator could stream out corresponding images and animation that could in turn portray the meaning of the lines in that language as also in the language the reader understands, thereby facilitating a naive, yet unassisted teaching system. PGL can also be used as a means for communication with the physically challenged.

(iii) As an aid to develop an animated multi-lingual dictionary:

PGL databases by themselves form a pictorial and animated version of a dictionary. Given such a set of databases for one language, generating the same for other languages can be achieved merely by adding the equivalent words in those languages.

(iv) As a cell phone application:

An instant universal translator application for mobile devices could be envisaged using PGL. With symbols grounded using the respective images and animations, effective communication especially for commonly required sentences such as: *I need to go to the airport*, *I need a glass of water*, *Where is the toilet?*, *Which train should I take?* could easily be queried with a person knowing only the local language. In countries where most of the local population comprehends only their national language, such a PGL based application can find great use in extracting information from the local person. The questions could be posed, for instance in English and corresponding answers in the national language could be translated back to English. Since the person who queries and the person who answers both crosscheck the semantics using PGL, the meaning is conveyed in a far more reliable manner. Coupled with a speech recognition and synthesis system, such a translator can prove to be a more powerful tool for both learning a new language and also for disseminating knowledge.

(v) As an add-on for Chat applications:

The messengers provided by Yahoo, Hotmail, Gmail, etc. facilitate exchanging of short conversations between two parties. The deciding factor for exchanging views in a chat application is the choice of a common language which both the parties are conversant in. PGL based machine-aided translation may alleviate the choice of a common language and conversations may be possible with parties knowing different languages.

4 PGL augmented Machine Translation System

In this section, we present our work on a PGL based translation system from a source language to a target language using a library of images and animations. We have considered the source language to be English and the target language to be Assamese, the official Indian language spoken in the north-eastern Indian state of Assam. The translation mechanism used in PGL based method is grossly different from those used in the rule-based or statistical machine translation methods. The architecture of the proposed translation technique is illustrated in Figure 3. The various blocks in the figure are described in the subsequent sections.

We treat the nouns and verbs within the input (source) text as labels. Using these labels a set of image databases is searched and the relevant image frames that portray them are pictorially rendered in a sequence or in an animated form so that they convey the actual meaning of the input text. Since these labels also have their corresponding equivalent words in the target language, a separate text string comprising these words of the target language is also generated alongside. The system thus renders the image frames along with the animations and the translated version of the text, thereby effecting translation. Care is taken to convert the target text into the proper form viz. Subject-Verb-Object or Subject-Object-Verb, as the case may be.

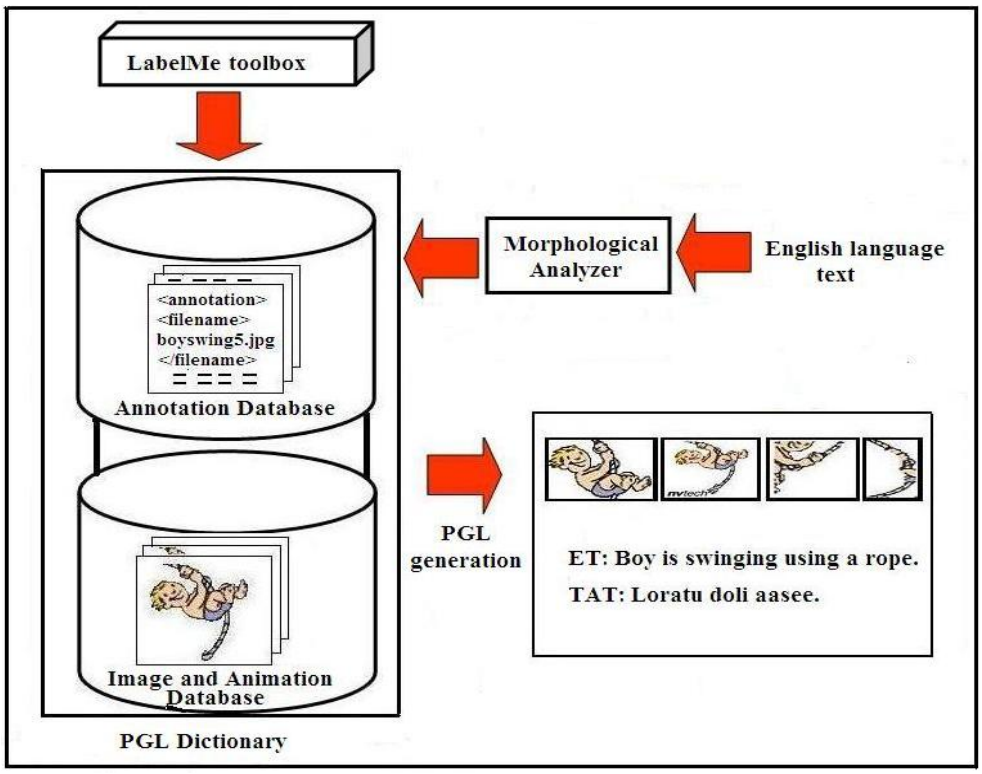


Figure 3. The Architecture of the PGL augmented Machine Translation System. ET: English Text, TAT: Transliterated Assamese Text

4.1 PGL Dictionary

The PGL dictionary, which is processed *a priori*, is made up of a collection of image frames extracted from several animated GIF files. The sequence of steps involved in its making is represented in the form of a flowchart as shown in Figure 4. Animated GIF files are a composition of image frames which are stitched together with a time delay introduced between successive frames. These image frames are then displayed one after the other recursively to render the animation. We have extracted each image frame from a set of animated GIF files using the freely available tool, *GIFFrame* (www.evanolds.com). We have used several animated GIF files depicting various actions for our initial translation test bed and stored them in the image and animation database.

We have used two freely available toolboxes namely *Annotation tool* (Korč 2007) and the *LabelMe toolbox* (Russell 2008) for image object annotation and searching. The annotations and their associated information (such as the source image frame, polygon points, class of the image objects, etc.) are stored as XML in files. Figure 5, depicts the use of the *Annotation tool* wherein the objects within the image of a boy swinging using a rope has been annotated. The objects within the image viz. the *boy* and the *rope* are separately selected and annotated (as *boy* and *rope*) and categorized (as *nouns*) in the respective languages (English and Assamese). This image frame thus serves to store two image objects viz. the *boy* and the *rope*. A separate image for a *rope* or a *boy* is thus not essential. Since sub-parts of every image are annotated, there is an efficient utilization of objects portrayed in every frame leading to a drastic reduction in the size of the PGL dictionary.

A sequence of image frames represents an action and falls under the verb class. These frames are assigned verb annotations such as *walk*, *run*, *play*, *swing*, etc. along with their equivalents in Assamese. If more languages need to be interpreted or translated, one needs to only add the corresponding annotations in that language. The system is thus scalable making the realization of a multi-lingual PGL augmented translation system comparatively easy. The dictionary also contains information required for morphological analysis.

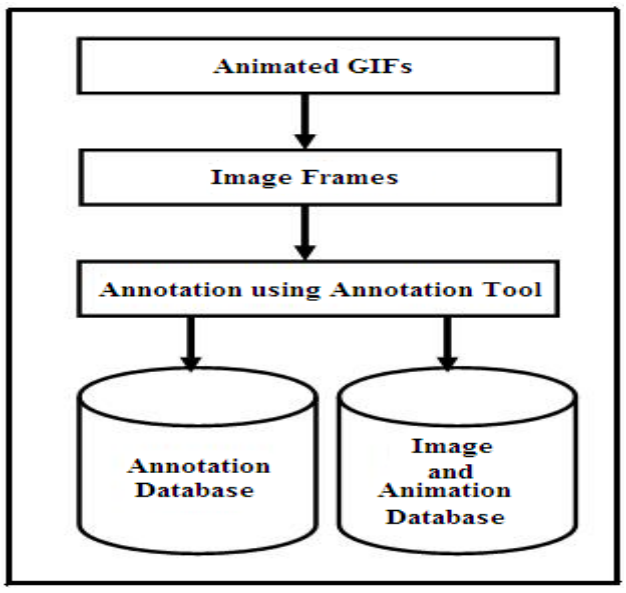


Figure 4. Flow chart for image pre-processing

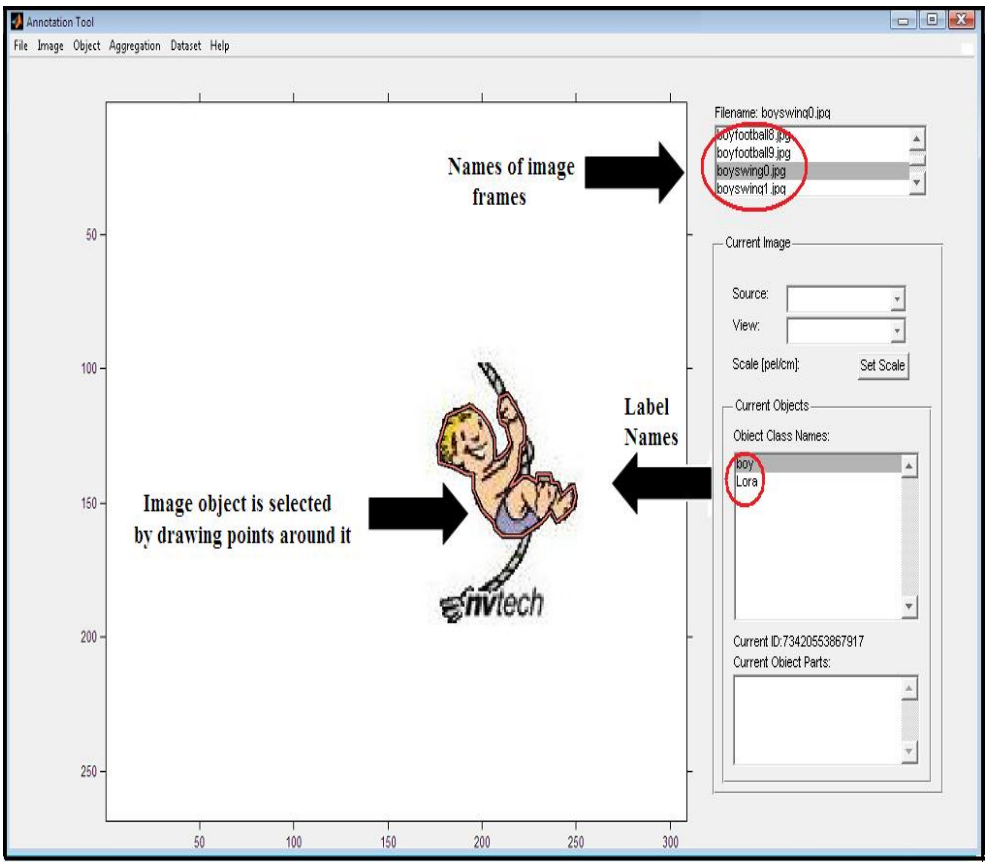


Figure 5. Image Object Annotation using the *Annotation toolbox*

4.2 PGL generation

The PGL generation phase consists of several steps which include morphological analysis, image/animation retrieval and translation from source text to the target text. The sequence of steps involved herein is depicted in the form of a flowchart in Figure 6.

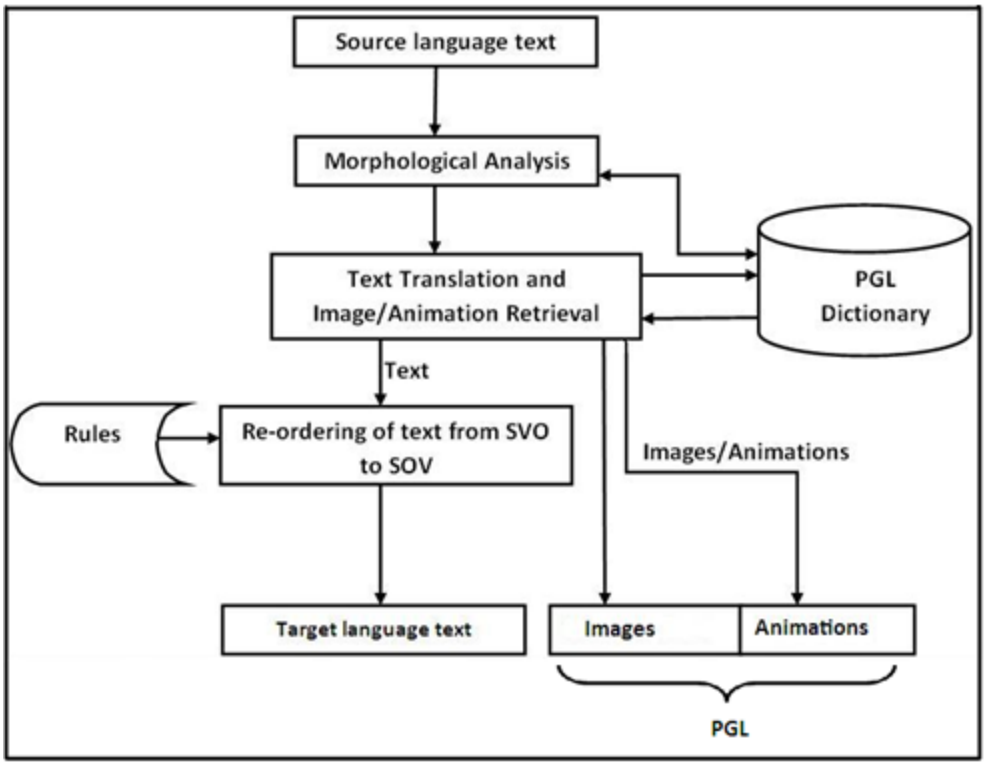


Figure 6. Flow chart representing the steps involved in PGL generation

4.2.1 Morphological Analysis

The role of the morphological analyzer is to apply morphological techniques over the given input source language text. The morphological analyzer takes into account a dictionary and a suffix list for performing morphological analysis. The analyzer makes use of stemming techniques (Jurafsky 2009). A given input text in the source language is first broken down into its constituent tokens. The tokens are verified using the morphological information within the PGL dictionary. The morphological component within this dictionary constitutes valid root words in the source and target languages. If a token is not a root word then a rule based suffix stripping technique is applied and the root word is derived and checked with the same dictionary. Table 2 shows part of the morphological information stored in the PGL dictionary. As the accuracy of translation by the system is dependent on morphological analysis, dictionary lookup and re-organizing according to the target language structure, it is necessary to enlarge the magnitude of the existing dictionary and the corresponding number of rules.

English root word	Assamese root word	Part of Speech	Transliterated English word	English Inflections	Assamese inflections	Transliterated English inflections
Boy	ল'ৰা	Noun	*Lora*	-s	-টো, -টোৰ, -বোৰ (	*-tu, -tur, -bur*
Girl	ছোৱালী	Noun	*Suwalee*	-s	-জনী, -বোৰ	*-janee, -bur*
Yellow	হালধীয়া	Noun	*saladhiyaa*	-ish	--	--
Dish	বাচন	Noun	*Basan*	-es	-বোৰ	*-bur*
Rope	ৰছী	Noun	*Rasi*	-s	-বোৰ	*-bur*
Cloth	কাপোৰ	Noun	*Kapur*	-s	-বোৰ	*-bur*
Walk	খোজ কাঢ়	Verb	*khuj kaar*	-ing	-কাঢ়ি	*- karee*
Play	খেল	Verb	*Khel*	-ing	-লি	*-li*
Swing	দোল	Verb	*Dol*	-ing	-লি	*-li*
Wash	ধুই	Verb	*Dhui*	-ing	-ছে	*-see*

Table 2. English/Assamese Morphological (partial) information stored within the PGL.

4.2.2 Image/Animation Retrieval

This module is responsible for searching and retrieving image objects and frames based on root words obtained after morphological analysis in the source language text. We have tailored the *LabelMe* Matlab toolbox to cater to our retrieval system. The *LabelMe* toolbox Matlab source code has, among many others, a rich set of functions for image object labeling, image manipulation, search tools and communication with the online annotation tool.

LabelMe tools for reading and plotting the annotations of individual files	
LMread	Reads one image and the corresponding annotation XML file
LMplot	Visualizes the image and polygons
LabelMe Database tools	
LMdatabase	Loads all the annotations into a big database structure
LMshowobjects	Shows the crops of the objects in the database
LMobjectnames	Returns a list with the name of all the objects in the database
LMobjectindex	Returns the indices of an object class within the annotation structure
LMcountobject	Counts the number of instances of an object class in every image
Search tools	
LMquery	Performs a query on the database using any field
LMobjectpolygon	Search for objects using shape matching
Image manipulation tools	
LMimread	Reads one image from the database
LMimscale	Scales the image and the corresponding annotation
LMimcrop	Crops the image and the corresponding annotation
LMcreateGIF†	Reads successive image frames and creates an animated GIF file.
* *Modified,* † *Newly added function*	

Table 3. Some pertinent functions used/modified/added in the *LabelMe* toolbox.

A few of the pertinent functions used in this toolbox along with their respective names and functions are listed in Table 3. Using the freely downloadable source code of this toolbox, we have augmented some of the existing functions with newer capabilities and also added new functions.

An image frame may contain a number of image objects, each of which is annotated separately by using the *Annotation* tool as mentioned earlier. The new functions of *LabelMe* incorporated in the toolbox are capable of extracting the desired image objects and also crop and display them as an image frame. This process of extracting objects from different images and then composing frames accordingly for rendering is performed on the fly. Finally, the set of frames are either rendered as static ones or as an animation by the animator.

4.2.3 Text Translation (English to Assamese)

This section discusses in detail the methodology of translation from English text to Assamese text. Assamese is a morphologically rich, agglutinative and relatively free word order language (Robinson 1839) (Saharia et. al. 2009). It has the *Subject-Object-Verb* (SOV) (Saharia et. al. 2009) grammatical structure unlike English which has *Subject-Verb-Object* (SVO) (Jurafsky 2009). Before conversion of the text from SVO to SOV, the system performs morphological analysis so as to improve the quality of the output. The conversion from SVO to SOV is achieved using a set of 20 linguistic rules stored within the PGL. The algorithm for this machine-aided translation follows the steps given below:

Step 1: Split up the English sentence into tokens.

Step 2: Look into the column heading "English root word" of the morphological content within the PGL dictionary with each of the tokens as the search key.

Step 3: If the token is inflectional, perform morphological analysis and derive the root word else goto Step 4.

Step 4: If the token is a root word and found in the PGL dictionary, perform steps 5 to 8.

Step 5: Identify the token whose part-of-speech is a verb from the column heading "Part of Speech" in the PGL dictionary. Rearrange the sentence into Subject (part before verb), Verb and Object (part after verb).

Step 6: Corresponding to the English root word, find its Assamese equivalent root word from the column "Assamese root word" from the PGL dictionary.

Step 7: The column "Assamese inflections" contains all the inflection forms of the Assamese root word. Add the suffixes to the Assamese root word for obtaining all the word forms in Assamese. Rearrangement of the words could be done by the users to refine and tune the semantic contents of the sentence.

Step 8: Rearrange, the Assamese words in SOV grammatical structure by using phrase tree (English) to phrase tree (Assamese).

We cite below how the translation algorithm works for two simple English sentences followed by a slightly complex sentences.

- ET†: The girl is smiling.
  ET† (After dictionary lookup/morphological analysis): <subject (The girl), verb (is smile)>.
  AT$^{¥}$(in SVO format): <subject (ছোৱালী), verb (আছে হাঁহি)> ।
  TET$^{€}$ (in SVO format): <subject (*Suwalee*), verb (*aasee hahee*)>.
  AT$^{¥}$(after adding suffixes and rearranging in SOV format): ছোৱালীজনীয়ে হাঁহি আছে ।
  TET$^{€}$ (after adding suffixes and rearranging in SOV format): *Suwaleejanee hahee aasee.*
- ET†: The girl is washing dishes with a cloth.
  ET† (After dictionary lookup/morphological analysis): <subject (The girl), verb (is wash), object (dish with a cloth)>.
  AT$^{¥}$(in SVO format): <subject (ছোৱালী), verb (আছে ধুই), object (কাপোৰ বাচন)> ।
  TET$^{€}$ (in SVO format): <subject (*Suwalee*), verb (*aasee dhui*), object (*kapur basan*)>.
  AT$^{¥}$(after adding suffixes and rearranging in SOV format): ছোৱালীজনীয়ে কাপোৰেৰে বাচনবোৰ ধুই আছে।
  TET$^{€}$ (after adding suffixes and rearranging in SOV format): *Suwaleejaneea kapuraree basanbur dhui aasee.*
- ET†: A boy is playing and another boy is fishing. (This sentence is a composition of two simple sentences connected by a conjunction).
  ET† (After dictionary lookup/morphological analysis): <subject (The boy), verb (is play)> and <subject (another boy), verb (is fish)>.
  AT$^{¥}$(in SVO format): <subject (এটা ল'ৰা), verb (আছে খেল)> আৰু <subject (আন এটা ল'ৰা), verb (আছে মাছ)>।
  TET$^{€}$ (in SVO format): <subject (*eta lora*), verb (*aasee khel*)> *aaru* <subject (*anya eta lora*), verb (*aasee mash)*>.
  AT$^{¥}$(after adding suffixes and rearranging in SOV format): এটা ল'ৰাই খেলি আছে আৰু আন এটা ল'ৰাই মাছ মাৰি আছে।
- TET$^{€}$ (after adding suffixes and rearranging in SOV format): *Eta lorai khelee aasee aaru aan eta lorai mash maaree aasee.*

†English Text, $^{¥}$Assamese Text, $^{€}$ Transliterated Assamese Text

4.2.4 Rendering the PGL

A Graphical User Interface allows the user to input a sentence in the source language. The morphological analyzer breaks the input into its constituent tokens, applies suffix stripping if necessary, verifies and passes each of the tokens one after another to the *LabelMe toolbox*. The tokens form the labels that we have used for annotating the image objects. The *LabelMe toolbox* accepts each of the tokens and queries the Annotation and Image databases. On finding the corresponding images, the *LabelMe toolbox* extracts the image object patterns from the databases and displays them to the user in the form of a sequence of image frames and animated GIF files based on whether the tokens were nouns or verbs. As mentioned earlier, the annotating of image objects is done both in English and Assamese. Thus, if we have annotated an image object with the English label "*boy*", the corresponding label in Assamese viz. "*lora*" is also provided. Thus the same image object is referenced by two different tokens viz. "*boy*" and "*lora*". This helps in translation from English to Assamese and vice versa. Having found both image objects and the equivalent annotations in the target language and sequenced them out in the desired order, the translated text is aligned in the

form of Subject-Verb-Object (SVO) or Subject-Object-Verb (SOV) as the case may be. While English uses the former, Assamese follows the latter. This reordering is performed as a post processing exercise before finally rendering the images and animations together with the translated text.

5 Results

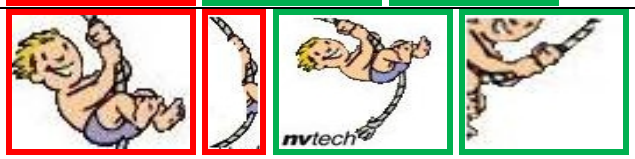
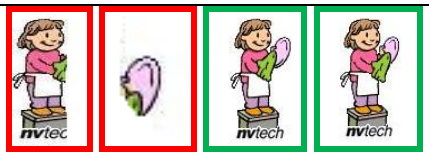
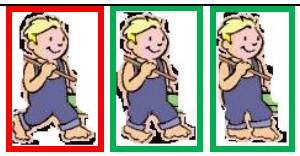
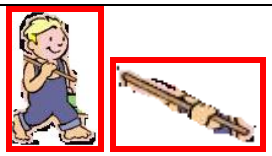
We have portrayed herein the results of our experiments based on a set of twelve animated GIF files shown in Figure 7(a). The 12 images shown are actually the first image frames of the associated animations. Figure 7(b) shows all the image frames of one such animation from Figure 7(a). The number of image frames per animation varies. Though this set of images may seem limited, there are altogether around 100 image frames constituting these animations. Further from each image frame we have extracted around 3 image objects which means the size of the database is around 300. The image objects are searched, extracted and rendered from within these image frames on the fly and no separate space exists for their storage.

The system was tested using several English sentences to be translated to their Assamese equivalent. Table 4 lists ten such example sentences along with the corresponding PGL comprising static noun images and animated verb image frames. For every sentence the corresponding objects and frames were accessed and a sequence of image frames and animated GIF files which depicted the related events in the sentence, were rendered, thus grounding the meaning of the sentence using the images and animations.

We also tested the system with a few slightly more complex English sentences like "*A boy is swinging using a rope and another boy is fishing with a fishing rod*", "*The boy is swinging using a rope and another boy is playing football*". We have observed that while the sentences are complex, they provide reasonably comprehensible PGL equivalents.

Figure 7(a). Animated GIF files used for testing (www.nvtech.com)

Figure 7(b). Image frames contained within one sample animated GIF file. The animated GIF image representing a boy is composed of 6 image frames

ET: The boy is swinging. AT: ল'ৰাটোৱে দোলি আছে। TAT: *Loratuya dolee aasee.*	
ET: The boy is swinging using a rope. AT: ল'ৰাটোৱে ৰছীৰ সহায়ত দোলি আছে। TAT: *Loratu rasir sahayat dolee aasee.*	
ET:The girl is washing dishes. AT: ছোৱালীজনীয়ে বাচনবোৰ ধুই আছে। TAT: *Suwaleejaneea basanbur dhui aasee.*	
ET:The girl is washing dishes with a cloth. AT: ছোৱালীজনীয়ে কাপোৰেৰে বাচনবোৰ ধুই আছে। TAT: *Suwaleejaneea kapuraree basanbur dhui aasee.*	
ET:The boy is walking. AT: ল'ৰাটোৱে খোজ কাঢ়ি আছে। TAT: *Loratuya khujkaree aasee.*	
ET:The boy has a fishing rod. AT: ল'ৰাটোৰ এটা বৰশী আছে। TAT: *Loratur eta barahee edaal aasee*	

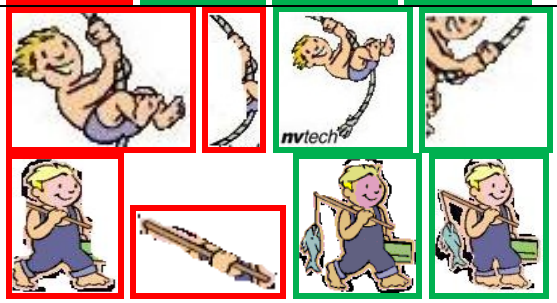
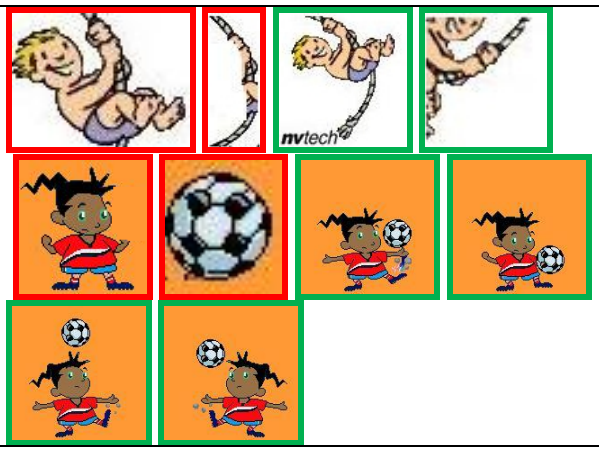
ET:The boy is playing football. AT: ল'ৰাটোৱে ফুটবল খেলি আছে। TAT: *Loratu football khelee aasee.*	
ET: The girl is running. AT: ছোৱালীজনীয়ে দৌৰি আছে। TAT: *Suwaleejaneea doiree aasee.*	
ET: A boy is swinging using a rope and another boy is fishing with a fishing rod. AT: এটা ল'ৰাই ৰছীৰ সহায়ত দোলি আছে আৰু আন এটা ল'ৰাই বৰশীৰে মাছ মাৰি আছে। TAT: *Eta loraai rasir sahayat dolee aasee aaru aan eta lorai baraheere mash maaree aasee.*	
ET: The boy is swinging using a rope and another boy is playing football. AT: ল'ৰাটোৱে ৰছীৰ সহায়ত দোলী আছে আৰু আনটো ল'ৰাই ফুটবল খেলি আছে। TAT: *Loratuya rasir sahayat doleei aasee aaru aantu lorai football khelee aasee.*	

Table 4: English Text (ET), its corresponding Assamese Text (AT), the Transliterated Text (TAT) and the static images (Nouns) and image frames constituting the animation rendered (Verb), all of which are generated on-the-fly. (Red bordered frames are static noun images while the green bordered ones are those that constitute the animations).

6 Conclusions

The use of image frames and animations that convey the meaning of a symbol in a language can be a vital aid in actually triggering the intrinsic grounding mechanism in a human being. Books that describe each of the words and sentences with appropriate illustrations tend to catalyze the language learning process as they ground each symbol in the language with its actual physical equivalent. In this paper we use this concept as a basis and describe an efficient method of machine translation that strives to provide a rendering of the semantics of the source language along with the equivalent target language, thereby grounding the symbols in both the source and target languages onto the same set of images and animations. The words (currently verbs and nouns) within the source language are used to carry out a search for relevant sets of image frames and objects within an *a priori* processed image and animation database. These are then sequenced and animated to render the required effect and convey the semantics of the source language sentence, along with the equivalent target language sentence. An English to Assamese machine-aided translation system was used to emphasize the effectiveness of the grounding of the meanings. Such a methodology can be used in conjunction with existing machine translation systems to avoid ambiguities that creep in the translation process. Since the images and objects within are tagged with both the source and target language equivalents, a reverse translation from target to source language will also produce the same sequence of image objects and animations. This will allow both the source language and the target language users to come to a consensus on the actual semantics being conveyed to one another, thereby improving the authenticity of the translation process. The system described can be scaled to any number of languages by tagging the images and the objects within with their corresponding equivalents in the new languages.

Like in most translation systems the initial effort in building the dictionary and corpora go unrecognized. Here too we wish to highlight the time and effort consumed in the *a priori* pre-processing of the GIF files as also the process of annotating image objects that comprise the PGL dictionary. However it may be noted that this is merely a one-time task, similar to the making of a conventional dictionary for a language. We are in the process of augmenting the dictionary with a variety of images so that a large range of image frames and objects facilitate better and effective translation.

We intend to use an Assamese morphological analyzer[1] so as to facilitate effective reverse translation from Assamese to English. This will allow for extraction of root words in the Assamese text which could in turn be used to search the relevant image frames and render them alongside the translated English text to finally provide for a PGL augmented bidirectional English-Assamese translator.

It may be noted that though the PGL system described herein does rely on the conventional machine translation system, it attempts to augment the output with vital grounding information required for the target language user. It may also be observed that the effectiveness of such a system is not completely revealed while translating a piece of literary text or poem. However it is likely that a simple story in one language could be correctly rendered to a child who comprehends another language. Another scenario where such PGL augmented translation can be effectively utilized is when for instance, an English speaking tourist needs to direct a Korean speaking cab driver to take him to the airport. Under such a situation the pictorially grounded rendering provided by this system would definitely ensure a faster and semantically reliable means of communication.

[1] (http://tdil.mit.gov.in/AssameseManipuri-IITGuwahatiJuly03.pdf)

7 Acknowledgements

We would like to thank Pallav K. Dutta, Indian Institute of Technology Guwahati, Assam, India and Akshay K. Vangari, Department of Computer Science, College of Engineering, California State University, Long Beach, CA, USA, for the assistance provided in verifying the authenticity of translations in Assamese and Spanish respectively. We also acknowledge the use of the GIF files created by NVTech Inc., Ottawa, Ontario, Canada, (www.nvtech.com) in our research work.

8 References

Alshawi, H., Bangalore, S., and Douglas, S., 2000, Learning dependency translation models as collections of finite state head transducers, *Computational Linguistics*, vol. 26, no. 1, pp. 45–60.

Bharati, A., Chaitanya, V., and Sangal, R., 1999, Natural Language Processing: A Paninian Perspective, Prentice Hall of India Pvt. Ltd.

Brown, P. F., Cocke, J., Pietra, S. D., Pietra, V. J. D., Jelinek, F., Lafferty, J. D., Mercer, R. L., and Roossin, P. S., 1990, A Statistical approach to Machine Translation, *Computational Linguistics*, vol. 16, no. 2, pp. 79–85.

Cangelosi, A., 1999, Modeling the evolution of communication: From stimulus associations to grounded symbolic associations, Proceedings of European Conference on Artificial Life, ECAL99, Springer-Verlag, pp. 654-663.

Chellamuthu, K. C., 2002, Russian to Tamil Machine Translation System at Tamil University, Proceedings of Tamil Internet 2002 Conference. (Available online: http://infitt.org/ti2002/papers/16CHELLA.pdf).

Chiang, D., 2007, Hierarchical phrase-based translation, *Computational Linguistics*, vol. 33, no. 2, pp. 201- 228.

Harnard, S., 1990, The Symbol Grounding Problem, *Physica*, D 42, pp. 345-346.

Hedden, D. T., 1992, Machine Translation: A Brief Introduction. (Available online: http://ice.he.net/~hedden//intro_mt.html).

Hutchins, W. J., 1986, Machine Translation: Past, Present, Future, Computers and their Applications Series, Ellis Horwood , Chichester.

Hutchins, W. J., 1995, Machine Translation: A Brief History, Concise History of the Language Sciences: from the Sumerians to the Cognitivists, E.F.K.Koerner and R.E.Asher (Eds.), Oxford: Pergamon Press, pp. 431-445.

Jurafsky, D., and Martin, J. H., 2009, Speech and Language Processing: An Introduction to Natural Language Processing, Speech Recognition, and Computational Linguistics, 2nd edition, Prentice-Hall.

Koehn, P., 2005, Europarl: A Parallel Corpus for Statistical Machine Translation, Proceedings of Machine Translation Summit, Thailand.

Koehn, P., Och, F. J., and Marcu, D., 2003, Statistical phrase-based translation, Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology, Canada, pp. 48–54.

Korč, F., Schneider, D., 2007, Annotation Tool. Technical report TR-IGG-P-2007-01, University of Bonn, Department of Photogrammetry.

Lewis, P. M. I., and Stearns, R. E., 1968, Syntax-directed transductions, *Journal of the ACM*, vol. 15, pp. 465–488.

Lopez, A., 2008, Statistical Machine Translation, *ACM Computing Surveys*, vol. 40, no. 3, article 8, pp. 1-49.

Robinson, W., 1839, A Grammar of the Assamese Language, Satyendra Narayan Goswami (Ed.), Purbadesh Mudran, Guwahati.

Russell, B. C., Torralba, A., Murphy, K. P., Freeman, W. T., 2008, LabelMe: A Database and Web-based tool for Image Annotation, *International Journal of Computer Vision*, vol. 77, nos. 1-3, pp. 157-173.

Saharia, N., Das, D., Sharma, U., and Kalita, J., 2009, Part of Speech Tagger for Assamese Text, Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP, Singapore, pp. 33-36.

Searle, J.R., 1980, Minds, brains and programs, *Behavioral and Brain Sciences*, vol. 3, no. 3, pp. 417-457.

Slocum, J., 1985, A Survey of Machine Translation: Its History, Current Status and Future Prospects, *Computational Linguistics*, vol. 11, no. 1, pp. 1-17.

Vogt, P., and Divina, F., 2007, Social Symbol Grounding and Language Evolution, *Interaction Studies*, vol. 8, no.1, pp. 31-52.

Wu, D., 1996, A Polynomial-time algorithm for Statistical Machine Translation, Proceedings of the Association for Computational Linguistics, ACL, pp. 152–158.