Building Multilingual Parallel Corpora from Scanned Pages of Digital Libraries and Gaming Techniques

Panasa Smita, K Kranthi, Rakesh Reddy T, Prahallad Lavanya

MSIT, International Institute of Information Technology, Gachibowli, Hyderabad, A.P 500032, India

Email: smita.panasa@gmail.com ; kkranthi438@gmail.com ; t.rakesh.reddy@gmail.com; lavanyap@andrew.cmu.edu

Abstract: There are a huge number of books that provide knowledge as well as entertainment to people. However these books cannot reach every person, as they are in languages that cannot be understood by them. It would be easier and a wonderful experience for readers if it were in their own native language. Several efforts were made to translate various books into Indian languages by a few people. However these efforts cannot be considered successful because manual process of translating books is tedious. Hence, to address this problem of translating books into various Indian languages we introduce Thingamacrib, a computer game that is enjoyable and which can be cherished always. People would love to play Thingamacrib because it is fun translating sentences. Players of Thingamacrib get entertainment, worthy gifts on playing the game and in addition to having a wonderful experience. The players not only help translating books into various Indian languages, but indirectly give joy to various others who can read their books in their native language.

INTRODUCTION

In the past few years, some of the famous books were translated into few Indian languages. However, this kind of effort lacks in translating technical books. There are many people who would like to gain knowledge on various fields. But their dreams are not fulfilled as most of the books are in English. In India alone, nearly 30% of the people do not know English. We believe that by providing books to people in their native languages, we can help people understand, gain knowledge and be more productive. High productivity leads to the development of the person and in turn leads to the nation's development.

This paper addresses the problem of translating books into Indian languages. We are motivated by the ESP game (which collects labels for images on the Web) and introduce the game Thingamacrib, which is a fun and exciting game of translating sentences. Thingamacrib's design ensures that the sentences translated through the game are correct. If our game is played as much as other popular games, like the ESP game, etc, we can translate a medium sized book in just a few weeks. Thingamacrib's design lays emphasis on creating a system

that is appealing and attractive to a large audience of people. This is in contrast to the previous approaches which relied on paid "experts" or "volunteers" or "writers". By introducing this game we have made the activity of translating books an enjoyable interactive process.

Thingamacrib is a two player game that helps in translating a sentence from one language to another language. There is no aspect of mentioning clues to one another in this game. To ensure 100% of the player's interest in the game, points are given to both players for each correct translated sentence. Based on the points earned, players win exciting gifts. By playing Thingamacrib, people help in translating books into various Indian languages not because they feel helpful, but as they have fun and earn wonderful gifts. The game can be played by anyone above 8 years of age.

Thingamacrib can be considered a "human algorithm", where a sentence in one language is given as input and a sentence in another language but with the same meaning as the input is given as output. This algorithm uses ordinary humans to interact with the computer throughout the Web instead of using a computer processor.

Thingamacrib has two parts to it: one is the synchronous game and the other is the asynchronous game. Both games are to be played online. Synchronous game requires two players who are teamed to play the game. This is the major part of Thingamacrib game. Asynchronous game can be played by a single player.

WHY A GAME?

Many people are not motivated towards translating books into various Indian languages. However, if they can be motivated, then many books can be translated. Many people find playing games as a means of relaxation in their busy schedules. Hence, we thought of developing a game that does multiple tasks together.

In addition to translating books, it provides relaxation and entertainment to the players. It also helps in keeping the players interested in the game. To achieve this, we have designed the game in an easy and interesting way. The players can easily understand the game and play it. Our game’s concept of providing points and gifts to players helps in developing a bond between players and the game.

SYNCHRONOUS GAME

OVERVIEW

Thingamacrib is meant to be played online by two players. Pairing of players is done randomly after their selection of languages. Each player needs to select two languages in which he/she wants to play the game.

Paired players have the same languages selected. One of the players is chosen as "Player1" and the other is "Player2". Player1 gets a sentence in the first language

selected. Player1 translates that sentence into the second language selected. This translated sentence is given to Player2. Player2 translates the sentence back into the first language selected.

THE GAME

For example, both players select *English* as first language and *Telugu* as the second language.

Player1 gets the following sentence as input:

"*rama gave book to sita*"

Player1 translates and submits the following output:

"rama sitaku pustakamu ichadu"

Player2 gets the following sentence as input:

"rama sitaku pustakamu ichadu"

Player2 now translates it back to English. If he enters the following sentence

"rama gave book to sita"

Player2's answer is compared with the Player1's input, i.e. both the English sentences are compared in this case. If they match, both players are given points. The translated sentence is stored in the database. For each sentence, each player gets a maximum of 45 seconds to translate it.

The player can skip a sentence if he/she feels it is difficult. One player is kept informed of the other player's status. For example, if Player2 skips a question, then Player2's status information given to Player1 is: "Player2 has skipped the question".

There is no limit of sentences to be played with. The game continues till one of the players quits the game. The other player need not loose heart as he can continue playing again with another player who is selected randomly.

No interaction is allowed between the players other than the sentences that are to be translated. The game also ensures that the sentence entered by a player is not same as the sentence received by that player for translation. There is no scope of hints in this game.

SCORING & RANKING

There is a timer which captures the amount of time both players took to translate their sentences. Based on this time, points are given. Here, time is inversely proportional to the points earned for translating a sentence. Both players earn the same number of points for a particular sentence. The lesser the time taken to translate, greater are the points received by the players.

In the version created, players get 5000 points each if both players translate in less than 30 seconds (together). If both players' sentences don't match, then they don't get any points for that sentence. Points are not subtracted for skipping a sentence, nor are incorrect answers penalized.

As in many other games, players get various ranks based on the scores they have achieved. Based on their ranks, exciting gifts are given to the players.

ASYNCHRONOUS GAME

THE GAME

Thingamacrib is usually meant to be played by two players. However, the asynchronous game can be played when the system cannot find a teammate for the player or if the player himself/herself wishes to play alone. Similar to the synchronous game, the player before starting the game needs to select two languages.

Player plays two sets of sentences. Each set consists of 5 questions each. Sentences which already have their translations in the selected languages are used in the first set. The second set consists of sentences which are not yet translated into the second language selected by the player. The second set of sentences may also be taken from the temporary files. The player gets a maximum of 2 minutes to translate the first set of sentences.

The player's proficiency is calculated based on his/her translations. Based on the player's proficiency, the system decides whether to give another first set of sentences or second set of sentences to the player. The player needs to get 60% or above proficiency to get the second set of sentences. The translations of the second set of sentences are stored in a temporary file. These sentences are given to four other players for translation. Sentences translated along with their translations are stored in the database once when all five players' translations match. There is no limit of sentences to be played with. The game continues till the player quits the game.

SCORING & RANKING

The player gets 7000 points for correct translation of the first set of sentences in 1 minute 30 seconds (considering all 5 sentences are translated correctly). Another 2500 points are given if the player's proficiency is above 60%. The player gets 500 points for translating the second set of sentences. The player has the option of skipping a sentence in both sets. However, scores would differ based on the number of correct translated sentences. If a sentence and its translation are moved from the temporary file to the database, 2500 points are added to all the five players automatically.

Points are not subtracted for skipping a sentence, nor are incorrect answers penalized. As in the synchronous game, players get various ranks based on the scores they have achieved. Based on their ranks, exciting gifts are given to the players.

SCREENSHOTS OF THE GAME

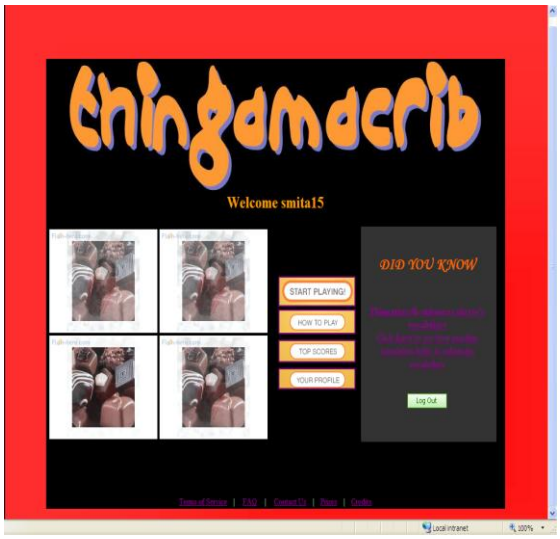

Fig 1: Screenshot of the website after a user has logged in

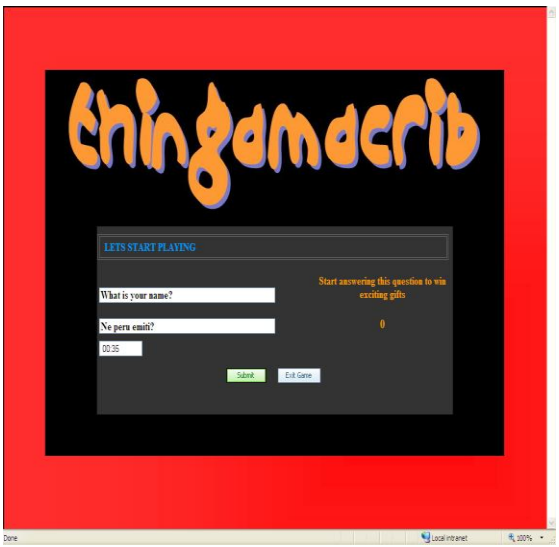

Fig 3: Screenshot of the game screen for the first player where he gets a sentence in English and translates it into Telugu

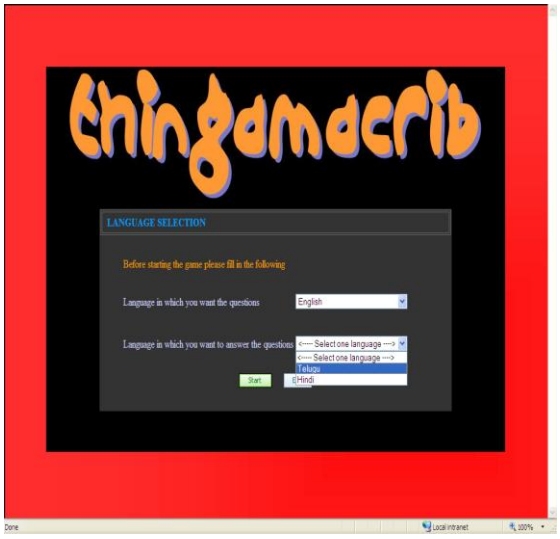

Fig 2: Screenshot of the Language selection page, where every player selects the two languages in which he/she would like o play

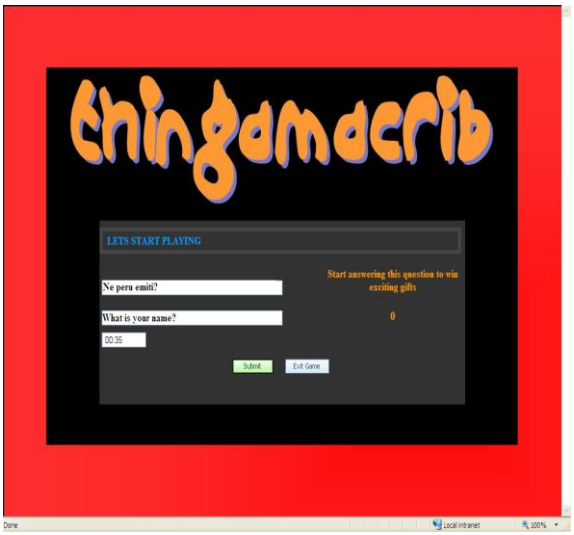

Fig 4: Screenshot of the game screen for the second player where he gets a sentence in Telugu and translates it into English

CONCLUSION

We have presented Thingamacrib, a game to translate books into various Indian languages. We have shown that Thingamacrib is enjoyable and fun to play. We have also shown how it helps to translate sentences correctly. The major contribution we present is the transformation of the tedious process of manually translating the books into various Indian languages by a few people into an enjoyable game which takes the help of many people from various parts of the world.

Just like the ESP Game, Thingamacrib can be considered as "human algorithm": humans act as processing nodes for problems that computers cannot solve yet. By providing exciting gifts as incentives for players, we gain a large quantity of computing power and data that can be used for multiple applications.

Translating the various books in the digital library into various Indian languages would be a classic achievement, and we believe Thingamacrib can be tremendously effective in doing so.

ACKNOWLEDGEMENT

Mr. Kishore Prahallad
Research Scientist, IIIT, Hyderabad, India
Mentor, MSIT, CIHL, Hyderabad, India
Mr. Vamshi Ambati
Mentor, MSIT, CIHL, Hyderabad, India

REFERENCES

Prahallad, L., Prahallad, K., Ganipathiraju, M., 2005. A simple approach for building transliteration editors for Indian languages. Proceedings from the International Conference on Universal Digital Library (ICUDL), Hangzhou, China. *Journal of Zhejiang University SCIENCE*, **6A**(11).

The Unicode Consortium, 2003. The Unicode Standard, Version 4.0. Addison-Wesley, Boston, MA.

The ESP Game: Labeling the Web, 2005. Carnegie Mellon University.