உத்தமம்
INFITT
உலகத் தமிழ்த் தகவல் தொழில்நுட்ப மன்றம்
International Forum for Information Technology in Tamil

Penn
University of Pennsylvania

மாநாட்டுக் கட்டுரைகள்
CONFERENCE PAPERS

TAMIL INTERNET 2011

தமிழ் இணையம் 2011

Agaraadhi: A Novel Online Dictionary Framework

Elanchezhiyan.K, Karthikeyan.S, T V Geetha,

Ranjani Parthasarathi & Madhan Karky

{chezhiyank@gmail.com, sethuramankarthikeyan@gmail.com, madhankarky@gmail.com}
Tamil Computing Lab (TaCoLa),
College of Engineering Guindy, Anna University, Chennai.

Abstract

This paper describes Agaraadhi, a dictionary framework for indexing and retrieving Tamil words, their meaning, analysis and related information. With a database of over 3 lack root words and their corresponding meaning in English and Tamil, this paper proposes a framework to encompass various features such as morphological analysis, morphological generation, word usage statistics, word pleasantness analysis, spell checking, similar word finder, word usage in literature, picture dictionary, number to text conversion, phonetic transliteration, live usage analysis from micro blogs and more. Describing various components of the framework the paper concludes with a discussion over dictionary statistics and possible features for future extension of the framework.

1. Introduction

Most of the Tamil dictionaries are synonym based and they do not give enough information such as morphological analysis of the word, possible case endings for requested word, pleasantness score, word usage in the web and social networks, equivalent words or meaning etc. To overcome these issues we propose Agaraadhi, a framework. Agaraadhi Framework consists of a Morphological analyser, Morphological generator, Word pleasantness and Word usage score finder as well as analysis of current usage in Social Networks, Picture dictionary, equivalent Tamil words, Generator (Word suggestions), Spell checker, Phonetic transliteration, Number to Text Converter, Rare-Word of the day and Social Network sharing.

Agaraadhi dictionary has more than 3 lac words in various domains such as General, Literature, Medical, Engineering, Computer Science, etc. The Agaraadhi framework dictionary is a Tamil English bilingual dictionary. The following sections describe the framework and list the benefits of such a framework over traditional online Tamil-English dictionaries. A few features proposed in this framework such as popularity score for a word, to best of our knowledge, are not present in any other world dictionaries.

2. Agaraadhi Framework

Agaraadhi Dictionary Framework was designed to provide additional information to the user regarding the word that they query about. Agaraadhi framework presented in figure 1 can be divided

into two major divisions, online and offline, in terms of the time of processing. This section describes the various components used in the Agaraadhi framework in detail.

2.1 Online Process

Any user query is sent sequentially to dictionary and literature, to retrieve corresponding data from those indices, fetching phonetic transliteration from transliteration modules, morphological information from morphological analysis and generator module and fetching live usage analysis from micro blogs. All those Information are sent to user interface pages, shown in fig 1.

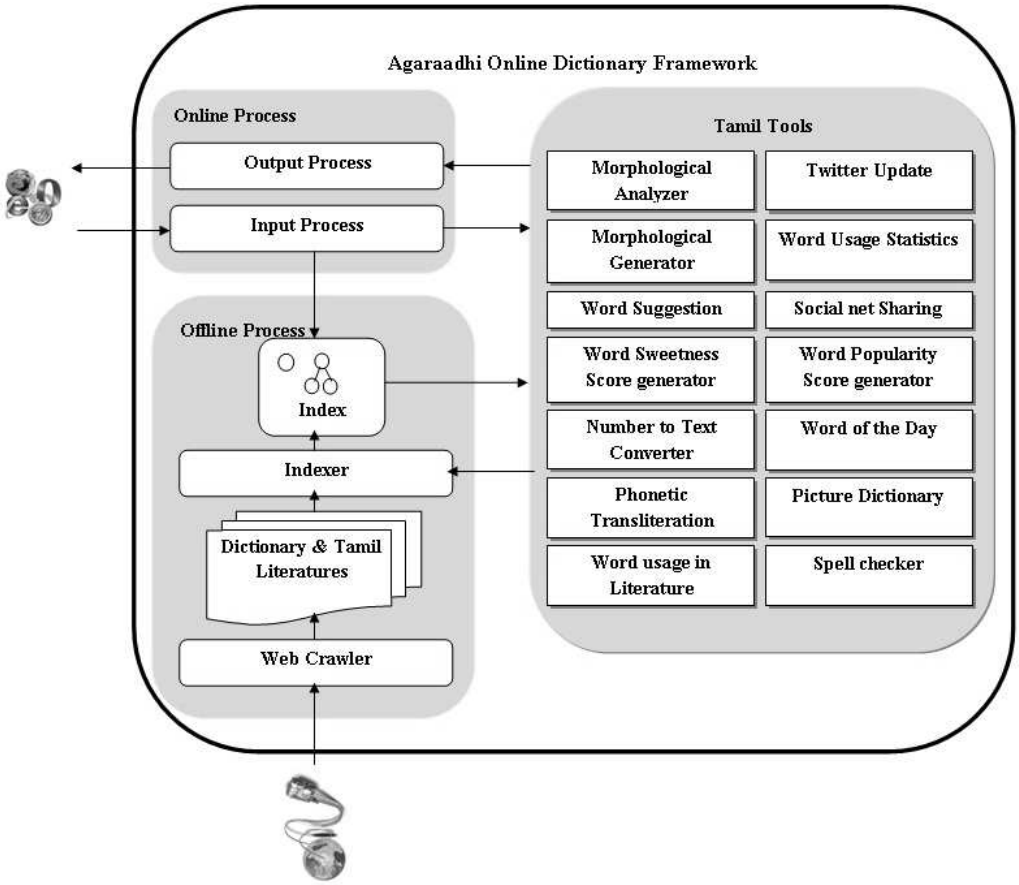

Fig 1: Agaraadhi Online Dictionary Framework

2.2 Offline Process

Tamil words and their meanings are entered manually and stored in text files. Those words are sequentially sent to modules such as popularity score generator, pleasantness score generator, picture dictionary, phonetic transliteration module and the resulting information is abstracted as a word object. Tamil literature such as Bharathiyaar songs, Avvaiyar songs, Thirukkural and lyrics are crawled using a static web crawler and are indexed in hash table as key value pairs.

2.3 Features of Agaraadhi Framework

Agaraadhi dictionary framework consists of more than twenty features such as Morphological Analysis Morph Generation, pleasantness scoring, popularity scoring, spell error suggestions etc.

3.1 Morphological Analyser

Morphological analyser [1] chunks the query word and gives the morphological features of the query word such as root word, parts of speech, gender, tense and count. If the Query word is *padithaan*, Morphological Analyser gives as *padi* as root, word represents male gender and query word is past tense and so on.

3.2 Morphological Generator

A Tamil morphological generator[2] needs to tackle different syntactic categories such as nouns, verbs, post positions, adjectives, adverbs etc. separately, since the addition of morphological constituents to each of these syntactic categories depends on different types of information. The generator is used to generate possible morphological variations of the query word.

3.3 Spell Checker

Spell Checker is used to check the spelling of Tamil words and to provide alternative suggestions for the wrong words. It uses the Morphological Analyzer. The Morphological Analyzer is used to split the given Tamil word into the root word and a set of suffixes. If the word is fully split by the analyzer and its root word is also found in the Agaraadhi dictionary, the given word is termed as correct. Otherwise, the correction process is invoked to generate all the possible suggestions with minimum variations from the given word.

3.4 Word Suggestions

Word Suggestion gives the list of equivalent or related words for the given query word.

3.5 Word Pleasantness and Word Popularity Score

Word Pleasantness score generator provides how easy to pronounce the word.

Word Popularity shows the word usage in the web. The Word from agaraadhi is given to web and found the frequency distribution of the word across the popular blogs, news articles, social nets etc.

3.6 Word Usage in Literature

This feature finds the usage of words in popular literature such as Thirukural, Bharathiyar Padalgal, Avvai songs and Lyrics.

3.7 Number to Text Converter

It converts a number to Tamil word equivalent as well as in English text. For example in Tamil we represent *oru Arpputham* (அற்புதம்) for 100 million, *Kumbam* (கும்பம்) **for** 10 billion and finally up to *Anniyan* (அந்நியம்) for one zillion.

3.8 Phonetic Transliteration

The pronunciation of words in Tamil and English language, as distinct from their written form based on the phonology and it can also vary greatly among dialects of a language. Phonetic transliteration module splits the word into syllables and gives the transliteration for each syllable.

3.9 Picture Dictionary

Pictures, photos or line drawings to depict popular words have been included in the dictionary to enable efficient learning for children using this tool.

3.10 Social net Sharing and Twitter Update

The framework also provides features to format results to be shared effectively on social networks. An Agaraadhi Bot was designed to post updates and word of the day on Twitter automatically.

3.11 Word of the Day and Word Usage statistics

A rare word is randomly chosen and is displayed in the opening page to facilitate users to learn a new word every day.

Word Usage Statistics [3] shows the usage of the word in the social network over the past one week.

3.12 Tamil Word Games

Games play a vital role in learning. Currently Agaraadhi has two Tamil word games namely Miruginajambo and Thookku Thookki. Miruginajambo is an unscramble game and Thookku Thookki is a Hangman game in Tamil.

4. Conclusion and Future Work

This paper describes Agaraadhi, an Online Dictionary Framework. Agaraadhi online dictionary is a bilingual dictionary containing over 3 lac words on various domains like General, Medical, Engineering, Computer science, Literature etc. This Online Dictionary framework encompass various features such as morphological analysis, morphological generation, word usage statistics, word pleasantness analysis, spell checking, similar word finder, word usage in literature, picture dictionary, number to text conversion, phonetic transliteration, live usage analysis from micro blogs etc. Providing APIs for programmers and developing mobile apps for Agaraadhi framework will open a good platform for many researchers and developers working in Tamil Computing area.

References

- Anandan, R. Parthasarathi, and Geetha, Morphological Analyser for Tamil. ICON 2002, 2002.
- Anandan, R. Parthasarathi, and Geetha, Morphological Generator for Tamil. Tamil Inayam, Malaysia, 2001.
- J. Jai Hari Raju, P. IndhuReka, Dr. Madhan Karky, Statistical Analysis and visualization of Tamil Usage in Live Text Streams, Tamil Internet Conference, Coimbatore, 2010.