Contextual spell checking for Tamil Language

Jananie Segar
Department of Computer Science
University of Jaffna
Sri Lanka

KengatharaiyerSarveswaran
Department of Computer Science
University of Jaffna
Sri Lanka

Abstract

A spell checker is a basic necessity for composing text in any language. However, for the language like Tamil, due to its complex nature, a little amount of reportable work have been done compare to languages like English and other Latin based languages. There are no fully functional spell checking systems available for Tamil language, especially as an open source software. This paper proposes a contextual spell correction feature to an existing spell checker that is developed using a hybrid approach by the authors. The existing spell checker is integrating the dictionary check, *Canti* check, crowd sourcing and suggestion generation. In addition to that a contextual spelling correction approach in which, mistakes that arise due to confusion in following sets of letters {*ழ akaram*, *ள akaram*, *lakaram*}, {*ந akaram*, *nakaram*, *ன akaram*} and {*ர akaram*, *rakaram*} is proposed in this paper. A bigram language model is used to make suggestions for the words with confusing letters. Performance of the proposed approach is 89.13% using a human evaluation.

Keywords:*Contextual Spell Checker, Tamil Language, Language model*

1.0 Introduction

A spell checker is a basic necessity for composing text in any language that find spelling mistakes and in many cases it provides most appropriate suggestion for the misspelled word. For the language like Tamil, due to its complex nature, relatively there is a little amount of work have been done successfully compare to languages like English and Latin based languages. There are several approaches used for spell checking and correcting mistakes. Dictionary, Rule-based, and n-gram are popular approaches. Though there are some proposals for Tamil spell checker [4][11][12],only few have been implemented[1][2][3][13]; however, none of them are fully functional and open source. This paper proposes a methodology for contextual Tamil spell checking on top of an existing spell checker developed using a hybrid approach.

1.1Challenges in Tamil spell checking

Agglutinative nature and the complexity in inflection make the Tamil a complex language to process using computers. Because, of this nature still there are no reportable language processing tools, including Tamil spell checkers, available for Tamil; specifically no fully functional prototypes can be found though there are few proposal.

Further, variations in written forms and poetic forms also make Tamil a complex language to process; the Colloquial variations, borrowed words and mixture of Sanskrit letters are also adding more toughness. For example, in a context, it will be difficult to tell whether a word is a correct word or not because of the above nature. Even some wrong word can be a correct word due to the context; may be in a story. For instance, *செல்றான்* is a correct word in the context of a dialogue in a short story, though it looks like a wrong word. Therefore, it will be difficult to develop a system by considering all these variations.

The *Canti* which comes at the joining of the two words in Tamil language also makes the processing hard. Omitting a joining letter or introducing a letter would change the meaning of the

word. For instance, *முக்கிய செய்திகள்* and *முக்கியச் செய்திகள்* give two different meanings because of the joining letter 'ச்'.
Relatively free order nature also another important attribute of Tamil that makes language processing difficult; especially when it comes to words that can act as nouns and verbs, such as *ஆண்டாள்*.

Finally, the ambiguity, which arises due to the confusion sets in Tamil also makes Spell checking is a complex task. Though Tamil scholars do not find difficulties in picking the right letter from following three sets {*ழ akaram, ள akaram, lakaram*}, {*ற akaram, nakaram, ன akaram*} or {*ர akaram, rakaram*} and use in their writing, other people find difficulties in picking the right one. This also becomes big problem due to the pronunciation of these letters. Variations in pronunciation in different regions also are adding complexity to spell checking. For instance, the letters '*ழ*' and '*ள*'are not correctly pronounced by the people in the Northern region of Sri Lanka; but those two are clearly pronounced by the people of Tamil Nadu in India. For instance if *பளம்*and *பழம்*are pronounced by a person in Northern Sri Lanka, person who listen and type those words may make mistakes. However, the spell check will not correct it if it uses only dictionary approach. For instance, in case of *அவன் பலம் சாப்பிட்டான்* (*Avaன palam cāppiṭ ṭ āன*) and *அவன் பழம் சாப்பிட்டான்* spell checkers with dictionary approach will not recognize the error. Here though both words *பலம்* and *பழம்* are correct according to dictionary, *பலம்* is wrong due to the context. Finding contextual errors are also a tough task in Tamil spell checking.

2.0 Literature survey

Dictionary approach, rule-based and n-grams are the common approaches used in existing researches. However, none of the efforts are integrating multiple approaches to improve performance of the spell checkers and to provide a complete solution.

An add-on for Mozilla Firefox called the *ThamiZha! Solthiruthi*[1] is a Tamil spell checker that correct misspelled words using dictionary approach. It maintains a Tamil corpus with close to 5 million words to correct spelling mistakes. The Indic Online spell checker [2] that also uses the dictionary approach and not a fully functional system which performs poor compare to [1].The *Nāvi*[3] is an Online *Cantipilaitritti* that corrects, *Canti*mistake, the join word error. This system checks whether there is a joining letter should come or not between two words. However, [3] checks only whether the existing *Canti* is an appropriate letter or not; it does not suggest a *Canti*letter if it is missing. There is an effort has been made to develop spell checker by using morphological analyser and make suggestions using Morphological generator [4] for Tamil. However, there are no systems found online for further study.

Few efforts have been made to correct contextual spelling mistakes using language models and error models [5] in other languages. An effort proposes spell checker that uses Machine learning to correct contextual errors [6]. In [6], to correct context dependent mistakes, confusion sets are prepared. Then a classifier is trained to choose appropriate letter from the confusion set by generating negative and positive examples. Thereafter it is used to correct contextual mistakes. In Microsoft Office2007, there is an improvement in using the contextual spell checking for English; which is shown using blue underscore for the words with contextual errors and the right click shows the appropriate suggestion[5]. However, this does not handle Tamil spelling mistakes and contextual mistakes. Another effort [7] on contextual spelling check for English language is proposed based on

n-gram technique. In which the most appropriate word is proposed based on the probability found in language model. For this purpose, the Google n-gram data set has been used [7].

3.0 Hybrid approach for Tamil spell checking

An application to check the Tamil spelling using a hybrid approach has been proposed and developed by the authors of this paper [8]. This approach integrates dictionary approach, *canti*check and crowd sourcing for new words. The *Levenshtein* distance finding algorithm [10] is used to match the words with the dictionary and flag the misspelled words. Grammatical rules have been written for *Valli̲ammikum* and *Valli̲ammikā* places to solve the *canti*problems. For generating suggestions, an n-gram based technique is used. Further a feature, called crowd sourcing, has been added to the system to collect new words from users. This is an important feature as there are a lot of colloquial words are used in Tamil. The proposed system shows an accuracy of 85.77 % using human evaluation [8].

4.0 Methodology

The hybrid approach proposed by the authors of this paper in [8] for Tamil spell checking does not handle contextual spelling mistakes. This paper proposes a method to correct the contextual spelling mistakes that occur due to the characters in the confusion sets {*ḻakaram, ḷakaram, lakaram*}, {*ṉakaram, nakaram, ṇakaram*} or {*ṟakaram, rakaram*} are considered. A bigram language model is used to mark errors and propose appropriate suggestions. The steps used in the proposed work is shown in figure 1.

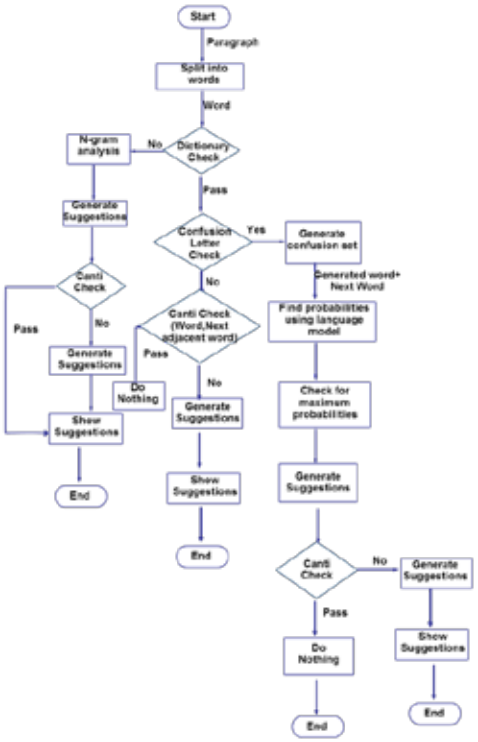

Figure 1: Proposed methodology

According to this approach, the given word is first checked with dictionary to see whether the word exists on the dictionary using *Levenshtein* algorithm. If it is not available, then the appropriate suggestions will be generated using letter level n-gram analysis. Then it is checked for joining letter, the *canti*letter. After the completion of these steps, word is checked to see whether there are any letter in confusion set are available in it. If such letter is available, then the letter is checked for appropriateness. This is done using a statistical approach with the aid of a bi-gram language model.

A corpus of Tamil text were collected from various sources, including the web. Then the text were cleaned to remove unnecessary characters and spaces. Next using the corpus, a bigram language model is prepared and probabilities for each pairs were calculated using an algorithm called Kneser-Ney smoothing without pruning [9]. A Language modelling tool called KenLM that is implemented based on the above algorithm is used to develop the bi-gram language model. Thereafter to accelerate the search, only the bi-gram pairs which have a confusing letter in it are extracted from the bi-gram language model.

For the words with a letter from confusion set, first the words will be generated with other letters in the same confusion set. For instance, if the given word is பலம் from the sentence

அவன்பலம்சாப்பிட்டான் then words *பழம்*, *பளம்* are also generated. After that these words will be checked with a dictionary to see whether they are available. For instance, in this case the words *பழம்*, *பளம்* will be checked with the dictionary. In this case,*பளம்* will not exist in the dictionary. Therefore, *பலம்* and *பழம்* are considered for further processing. Then the following word to the *பலம்* is also extracted. After that both words are checked for probabilities in language model. According to this example, it will be *பலம்சாப்பிட்டான்* and *பழம்சாப்பிட்டான்*. Finally, based on the probabilities found in the bi-gram language model most appropriate word (whether *பலம்*or*பழம்*) will be suggested.

5.0 Results and Discussions

This concept has been developed as an online tool. In this tool the different types of mistakes are marked using different colours as follow:

- Mistakes after dictionary check are shown in red colour
- Mistakes due to the contextual spell checking are shown in green colour
- The correct words will be shown in black colour
- *Canti* mistakes are shown by making words bold

The suggestions that are generated for the mistakes are shown to users. 'Add to dictionary' link is also shown to users as an option. If user clicks the 'add to dictionary', then the word will be sent to the consideration of administrator who will then validate and add to the dictionary.

This proposed approach shows an improvement over the earlier work proposed in [8] by the authors of this paper. An accuracy of 89.13% was obtained for human evaluation, which is promising compare to the previous effort. However, the success of this contextual spell checking is heavily depend on the size of the corpus. If the corpus has all possible words and the combinations for confusion letters, then high accuracy can be achieved.

Unavailability of global testing data set for Tamil spell checkers is a problem. Because of this the evaluation made by the researchers are becoming subjective. Therefore, the test data that is used in this research is published along with the application. Further, the authors of this paper are researching to publish a standard test data for spell checking of Tamil language by considering all possible errors that can occur in Tamil language.

6.0 Conclusion

The proposed approach shows promising results compare to previous efforts. Hybrid approach for Tamil spell checking with the contextual spell correction feature to correct {*ழ akaram, ள akaram, lakaram*}, {*ந akaram, nakaram, ன akaram*} and {*ர akaram, rakaram*} errors clearly improve the accuracy of the spell checker for Tamil language; the accuracy of 89.13% has been obtained. The proposed work has been implemented and hosted in thamizhi.org/akaram.

References:

[1]. Ve. Elanjelian, S. Muguntharaj, Radhakrishnan, Vijay, A. Suji, MalathiSelvaraj, Sri Ramadoss, YagnaKalyanaraman, LavanyaSundararajan, PranavaSwaroop and SiddarthVengatesan,"ThamiZha!Solthiruthi(Tamilspellchecker)",Internet:https://addons.mozilla.org/en-US/firefox/addon/thamizhasolthiruthi/?src=search,February 04,2012.

[2]. SanthoshThottingal,"Indic Online Spellchecker", Internet**:** http://thottingal.in/projects/ spellchecker/[Nov.15, 2013].

[3]. "*நாவி*", Internet**:** http://dev.neechalkaran.com/p/naavi.html#.VNjoIubF8hY[Oct.20, 2013].

[4]. Dhanabalan,T., RanjaniParthasarathi, and T. V. Geetha. "Tamil Spell Checker." Sixth Tamil Internet 2003 Conference, Chennai, Tamilnadu, India. 2003.

[5]. "Contextualspellinginthe2007Microsoftofficesystem", Internet: http://blogs.msdn.com/b/correcteurorthographiqueoffice/archive/2006/06/05/ contextual-spelling-in-the-2007-microsoft-office-system.aspx, June.05, 2006. [Jan.05, 2015].

[6]. AllaRozovskaya and Dan Roth, "Generating Confusion Sets for Context-Sensitive Error Correction",EMNLP '10 Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing,pp. 187-197.

[7]. Carlson, Andrew, and Ian Fette. "Memory-based context-sensitive spelling correction at web scale." Machine Learning and Applications, 2007. ICMLA 2007. Sixth International Conference on. IEEE, 2007.

[8]. S.Jananie,K.Sarveswaran, "HYBRID APPROACH FOR SPELL CHECKING OF TAMIL LANGUAGE", International Research Sessions, University of Peradeniya, Sri Lanka, Vol. 18, 4th & 5th July, 2014,pp.900.

[9]. Heafield, Kenneth. "KenLM: Faster and smaller language model queries."Proceedings of the Sixth Workshop on Statistical Machine Translation. Association for Computational Linguistics, 2011.

[10]. Natural Language Processing, Topic: "Minimum Edit Distance", Internet: https://web. stan ford.edu/ class /cs124/lec/ med.pdf, Stanford University.

[11]. *ம.பார்கவிகி*, *உமாதேவி*,"*தானியங்குசொற்பிழைதிருத்திஉருவாக்கத்தில்உருபன்பகுப்பியின்பங்கு*",Tamil Internet Conference, June 23-27, 2010.

[12]. R.*பத்மமாலா*,*ம.பார்கவி*,"*தானியங்குசந்திப்பிழைதிருத்திஉருவாக்கத்தில்உருபன்பகுப்பியின்பங்கு*", Tamil Internet Conference, June 23-27, 2010.

[13]. AnithaS.Pillai," Spell Checker for Tamil using Finite State Automata,Tamil Internet Conference, June 23-27, 2010.