Detecting and Correcting Contextual Mistakes in Sentences Using Part of Speech Tags

ABSTRACT

- A grammar checker is a tool to check each sentence in a text to see whether it conforms to the grammar. In case it finds a structure that conflicts with the conformity to the grammar, it would give suggestions for alternatives. The grammar checkers for European languages and some Indic languages are well developed. However, perhaps, owing to Tamil being a morphologically rich and agglutinative language this has been a challenging task. An approach to detecting and correcting grammatical mistakes due to subject and finite-verb disagreement with regard to person, number and/or gender and due to disagreement in tense aspects in Tamil sentences is proposed in this paper. A method has been proposed that uses hierarchical part-of-speech tags of words to detect the grammatical mistakes in subject and finite-verb agreement and mistakes in tense aspects in Tamil sentences. Two sets of Tamil grammar rules are used to generate suggestions for the grammatical mistakes. Test results show that the proposed grammatical mistake detection and correction system performs well.

- *Keywords: Tamil, Part-of-Speech, Grammar, Subject, Suggestion*

1. INTRODUCTION

Tamil, a Dravidian language and an official language of Sri Lanka, is an agglutinative language having rich morphological structure. Tamil has 247 letters comprised of 12 vowels, 18 consonants, 216 composite letters combining each consonant with each vowel, and one special letter, known as “ayutha eluththu” [1] [2] [3]. Tholkappiyam [4] and Nannool [5] are two popular works in Tamil grammar that define how words and sentences have to be formed in Tamil. Tamil words are formed by lexical roots followed by one or more suffixes. Tamil verbs inflect with *person*, *number* and *gender marking*, and combine auxiliaries that indicate *aspect*, *mood*, *causation* and *attitude* whereas Tamil nouns inflect with *plural markers*, *case markers* and *gender markers* [1] [2] [3]. The agglutinative nature makes Tamil a complex language to process using computers.

The grammatical error detection and correction process can be divided into two types namely *interactive* and *automatic*. Interactive grammar checkers would find the grammatical errors and suggest the most suitable alternatives to the grammatical mistakes, allowing users to select the best one from the suggestions given or ignore the suggestions. Automatic grammar checkers would find the grammatical errors and automatically correct them by substituting the most suitable suggestion in context, without user intervention.

Everyday humans produce several research articles, books, documents and magazines in this universe. When typing these documents, several grammatical errors may occur in sentences, such as *subject-verb disagreement*, *inconsistent tense aspects* and *modifier-noun disagreement* [6] [7]. These errors can be detected and corrected with suitable replacements by considering the syntactic categories (part-of-speech tags of words) of each word in the sentence. Humans may take a long time to find and correct errors in the whole document or fail to detect all of them due to ignorance and tiredness. An efficient grammar checker can help detect and correct grammatical errors quickly without being tired and becomes as an essential tool for text processing applications.

In this paper, we focus on detecting and correcting mistakes in *subject and finite-verb agreement* with regard to *person*, *number* and/or *gender* and mistakes in *tense aspect agreement* in Tamil sentences and analysing the necessity of part of speech tags of words to handle these grammatical mistakes. Tamil words to develop a grammar checker.. Hierarchical part of speech (POS) tags of words are used to detect these types of grammatical mistakes, and *two* sets of Tamil grammar rules are used to generate suggestions.

After this introductory section, the rest of this paper is organised as follows: Section II describes our methodology. Test results are discussed in Section III, with the conclusion in Section IV.

2. METHODOLOGY

The system proposed in this paper aims to detect and correct the following types of grammatical mistakes in Tamil sentences.

Type 1: Mistakes in subject and finite-verb agreement with regard to person, number and/or gender

In Tamil sentences, the finite-verb must agree with the subject of the sentence in terms of person (first person, second person, third person), number (singular, plural) and gender (masculine, feminine). For example, in the sentence ‘*அவன் நேற்று பாடசாலைக்குச் சென்றாள்*’, subject and finite-verb disagree in gender (‘*அவன்*’ is of masculine but ‘*சென்றாள்*’ is of feminine). In the sentence ‘*அவர் மைதானத்தில்*

விளையாடினார்கள்', subject and finite-verb disagree in number ('அவர்' is singular but 'விளையாடினார்கள்' is plural).

Type 2: Mistakes in tense aspect agreement

In Tamil sentences, the words that indicate the tense (past tense, present tense or future tense) must agree. For example, in the sentence 'அவன் நேற்று பாடசாலைக்குப் போவான்', the word 'நேற்று' indicates past tense but the finite-verb 'போவான்' indicates future tense causing disagreement in tense aspect.

Hierarchical part-of-speech (POS) tags of words are used to detect the grammatical mistakes and two sets of Tamil grammar rules are used to generate suggestions. The first set of Tamil grammar rules is used to generate suggestions to correct mistakes due to subject and finite-verb disagreement, and the other one is used to generate suggestions for tense-aspect disagreement. These grammar rules have been defined in the form of Python conditional statements based on standard Tamil grammar.

To determine the hierarchical POS tags, the POS annotated corpus provided by the Computational Linguistic Research Group (CLRG), AU-KBC Research Centre, MIT campus of Anna University [8] was used. Their tag-set is of hierarchical category having four levels consisting of 47 labels. Words with tags PR-PRP, N-NN and V-VM-VF are considered for assigning sub-tags. Here, PR-PRP indicates personal pronoun, N-NN indicates common noun and V-VM-VF indicates finite-verb and these are the essential parts for grammar check. The sub tags, namely, FP, SP, TP, SL, PL, M, F, PA, PR, FU, CM2, CM3, CM4, CM5, CM6 and CM7 that indicates indicate *first person*, *second person*, *third person*, *singular number*, *plural number*, *masculine gender*, *feminine gender*, *past tense*, *present tense*, *future tense*, *accusative*, *instrumental*, dative, *ablative*, *genitive* and *locative* case markers respectively are not assigned to words in AU-KBC annotated corpus. The hierarchical POS tags together with the sub-tags of words in the sentences are determined using a hierarchical POS tagger [9].

Now let us discuss how to detect and correct these grammatical mistakes in Tamil sentences.

2.1 DETECTING AND CORRECTING GRAMMATICAL MISTAKES OF TYPE 1

For a Tamil sentence, its finite-verb plays a major role as it shows tense, number, gender and person of the subject, even if the subject is not explicitly mentioned. It also demonstrates the functions of the subject. The finite-verbs are identified in the sentences using the POS tag V_VM-VF. The subject of a sentence can be a common noun or proper noun that must agree with the finite-verb in terms of person, number and gender. In order to check this agreement, we have to identify the subject and finite-verb.

For example, for this input sentence 'அவள் கோயிலுக்குப் போனாள்', the tag sequence will be {PR-PRP-SL-F-TP, N-NN-SL-CM4, V-VM-VF-PA-SL-F-TP} respectively. In the third tag, VF indicates *finite-verb* and sub-tag F of it indicates *feminine gender*, SL indicates *singular number*, TP indicates the person as a *third person* and PA indicates *tense past*. In N-NN-SL-CM4, N indicates *noun*, NN indicates *common noun*, SL indicates *singular number* and CM4 indicates *dative case marker*. Because of case marker, 'கோயிலுக்குப்' cannot be considered as a subject. In PR-PRP-SL-F-TP, PR indicates *pronoun*, PR indicates *personal pronoun*, F indicates *feminine gender*, SL indicates *singular number* and TP indicates person as a *third person*. The corresponding word 'அவள்' will be taken as a subject. Now we have to check the finite-verb tag and noun tag to see whether these tags agree in terms of *person*, *number* and *gender*. In both tags, gender indicator (F) agrees, number indicator (SL) agrees and person indicator (TP) agrees.

There may be cases any one or more of these in disagreement. For example,

Example 1:

அவன்	கோயிலுக்குப்	போனாள்
PR-PRP-SL-M-TP	N-NN-SL-CM4	V-VM-VF-PA-SL-F-TP

In the tag PR-PRP-SL-M-TP of the subject, the sub-tag M indicates masculine gender whereas the sub-tag F in the tag V-VM-VF-PA-SL-F-TP of the finite-verb indicates feminine gender. This conflict causes a grammatical mistake that should be corrected.

Example 2:

நான்	கோயிலுக்குப்	போனான்
PR-PRP-SL-FP	N-NN-SL-CM4	V-VM-VF-PA-SL-M-TP

In the tag PR-PRP-SL-FP of the subject, the sub-tag FP indicates first person whereas the sub tag TP in the tag V-VM-VF-PA-SL-M-TP of the finite-verb indicates third person. This conflict causes a grammatical mistake that should be corrected.

Example 3:

நான்	கோயிலுக்குப்	போனோம்

PR-PRP-SL-FP N-NN-SL-CM4 V-VM-VF-PA-PL-FP

In the tag PR-PRP-SL-FP of the subject, the sub-tag SL indicates singular number whereas the sub-tag PL in the tag V-VM-VF-PA-PL-FP of the finite-verb indicates plural number. This conflict causes a grammatical mistake that should be corrected.

These mistakes can be corrected by changing the corresponding suffixes that indicate person, number and gender. In Example 1 above, the sub-tag (M) of the subject indicates masculine gender whereas the sub-tag (F) of the finite-verb indicates feminine gender. The sub-tag M corresponds to the suffix '*அன்*' in the subject '*அவன்*' whereas F corresponds to the suffix '*ஆள்*' in the finite-verb '*போனாள்*'. In order to remove the disagreement, the sub-tag corresponding to gender of the subject should be F or the sub-tag corresponding to gender of the finite-verb should be M, and thus the change in corresponding tags would remove the disagreement. In other words, the subject can be '*அவள்*' while keeping the finite-verb as given or the finite-verb can be '*போனான்*' while keeping the subject as given. Similarly, in Example 2, in order to remove the disagreement, the sub-tag corresponding to person of the subject should be TP or the sub-tag corresponding to person of the finite-verb should be FP. Thus the subject can be '*அவன்*' while keeping the finite-verb as given or the finite-verb can be '*போனேன்*' while keeping the subject as given. In Example 3, to remove the disagreement, the sub-tag corresponding to the number of the subject should be PL or the sub tag corresponding to the number of the finite-verb should be SL, and thus the subject can be '*நாம்*' while keeping the finite-verb as given or the finite-verb can be '*போனேன்*' while keeping the subject as given.

However, alternative suggestion for subject while keeping finite-verb is given for personal pronouns of third person, namely, {'*அவன்*', '*அவள்*', '*அவர்*', '*அவர்கள்*'}. For other subjects, alternative suggestion is given for finite-verb only. For example, for the erroneous sentence '*ஆசிரியர் போனான்*', the suggestion is given only for the finite-verb, not for the subject. Another kind of errors that can be detected but giving suggestions is not considered because of the unavailability of a good morphological analyser. For example, for the erroneous sentence '*மாடுகள் விரைவாகப் ஓடிப் போனது*', subject and finite-verb disagrees in number. However, giving suggestion for this type of error is not considered. In this example, '*போனது*' should be changed into '*போயின*'. A good morphological analyser is essential in order to generate the suggestion to this type of finite-verbs.

2.2 DETECTING AND CORRECTING GRAMMATICAL MISTAKES OF TYPE 2

Disagreement in tense aspects causes mistakes of Type 2. Finite-verbs that have sub-tags PA, PR and FU (indicating past, present and future, respectively) are words of tense aspect. Another set of words that indicate tense are also identified. For example, words like '*நேற்று*', '*கடந்த*' indicate past tense, words like '*இன்று*', '*இன்றைய*' indicate present tense and words like '*நாளை*', '*அடுத்த*' indicate future tense. The words that have tense aspects in a sentence must be of the same tense. If a disagreement is found, the word that disagrees in the tense aspect should be changed suitably.

Let us consider a sentence to see how a mistake in tense agreement can be detected: '*அவள் நேற்று பாடசாலைக்குப் போவாள்*'. The POS tagger assigns tags to each word in this sentence. The tag sequence (including sub-tags) of this sentence is {PR-PRP-SL-F-TP, N-NN-PA, N NN-SL-CM4, V-VM-VF-FU-SL-F-TP}. Here PA indicates the tense aspect in word '*நேற்று*' and FU indicates the tense aspect in word '*போவாள்*'. PA indicates past while FU indicates future causing a disagreement. In this case, either the word '*நேற்று*' of past tense should be replaced by a word to indicate future tense while keeping the finite-verb as given, or the tense marker '*வ்*' of the word '*போவாள்*' indicates future should be replaced by tense marker '*ன்*' to indicate past tense without changing word '*நேற்று*'. If the finite-verb is kept as given '*நேற்று*' should be changed by a word to reflect future tense. Suggestion may be given from the set of words {'*நாளை*', '*அடுத்த வாரம்/மாதம்/வருடம்*', '*வருகின்ற வாரம்/மாதம்/வருடம்*'}.

The flowchart for detecting and correcting mistakes in subject and finite-verb agreement, and mistakes in tense aspect agreement in Tamil sentences is projected in Fig. 1.

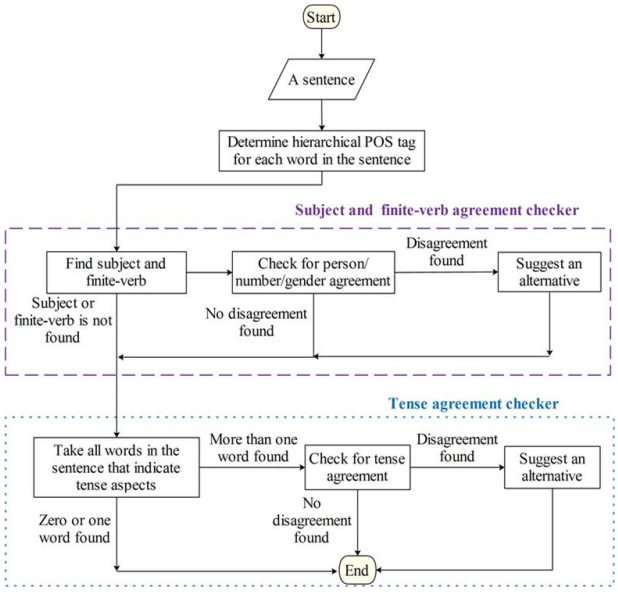

Fig. 1: Flowchart for correcting mistakes in subject and finite-verb agreement, and mistakes in tense aspect agreement in Tamil sentences

3 RESULTS AND DISCUSSION

During the process of grammatical error detection and correction, two entities are recorded:

1. words that cause grammatical mistakes and
2. generated possible alternatives for them.

The following four examples show the grammatical error detection and correction output.

Example 1: Sentence with mistake in gender agreement

Input sentence: *அவள் நேற்றுச் சந்தைக்குப் போய் நிறையப் பழங்களை வாங்கிச் சாப்பிட்டான்.*

POS tagger output: *அவள்*/PR-PRP-F-TP-SL *நேற்றுச்*/N-NN-PA *சந்தைக்குப்*/N-NN-SL-CM4 *போய்*/V-VM-VNF-VBN *நிறையப்*/JJ *பழங்களை*/N-NN-PL-CM2 *வாங்கிச்*/V-VM-VNF-VBN *சாப்பிட்டான்*/V-VM-VF-M-TP-SL-PA ./RD-PUNC

Subject: *அவள்*

Finite-verb: *சாப்பிட்டான்*

Subject and finite-verb disagree in gender (*அவள்* is of feminine gender but *சாப்பிட்டான்* is of masculine gender).

Suggestions: (*அவன்*/*சாப்பிட்டான்*) or (*அவள்*/*சாப்பிட்டாள்*)

In this input sentence, the grammatical error detection system identifies the word '*அவள்*' as a subject and the word '*சாப்பிட்டான்*' as a finite-verb based on the main POS tags. The sub-tag F of the subject indicates feminine gender whereas the sub-tag M of the finite-verb indicates masculine gender causing the disagreement in gender. In order to remove this disagreement, the correction system gives (*அவன்*/*சாப்பிட்டான்*) or (*அவள்*/*சாப்பிட்டாள்*) as suggestions based on the POS gender sub-tags.

Example 2: Sentence with mistake in number agreement

Input sentence: *அவர் காலையில் எழுந்தவுடன் முகத்தை சுத்தமான நீரினால் நன்றாகக் கழுவினார்கள்.*

POS tagger output: *அவர்*/PR-PRP-TP-SL *காலையில்*/N-NN-SL-CM2 *எழுந்தவுடன்*/V-VM-VNF-RP *முகத்தை*/N-NN-SL-CM2 *சுத்தமான*/JJ *நீரினால்*/N-NN-SL-CM3 *நன்றாகக்*/RB *கழுவினார்கள்*/V-VM-VF-PL-TP-PA ./RD-PUNC

Subject: *அவர்*

Finite-verb: *கழுவினார்கள்*

Subject and finite-verb disagree in number (*அவர்* is singular number but *கழுவினார்கள்* is plural number).

Suggestions: (*அவர்கள்*/*கழுவினார்கள்*) or (*அவர்*/*கழுவினார்*)

In this input sentence, the grammatical error detection system identifies the word '*அவர்*' as a subject and the word '*கழுவினார்கள்*' as a finite-verb based on the main POS tags. The sub tag SL of the subject indicates singular number whereas the sub-tag PL of the finite-verb indicates plural number causing the disagreement in number. In order to remove this disagreement, the correction system gives (*அவர்கள்*/*கழுவினார்கள்*) or (*அவர்*/*கழுவினார்*) as suggestions based on the POS number sub-tags.

Example 3: Sentence with mistake in person agreement

Input sentence: நான் நேற்று திரையரங்கிற்கு சென்று புதிதாக வெளிவந்துள்ள படத்தைப் பார்த்தான்.
POS tagger output: நான்/PR-PRP-FP-SL நேற்று/N-NN-PA திரையரங்கிற்கு/N-NN-CM3 சென்று/V-VM-VNF-VBN புதிதாக/RB வெளிவந்துள்ள/V-VM-VNF-VBN படத்தைப்/N-NN-CM2 பார்த்தான்/V-VM-VF-M-TP-SL-PA ./RD-PUNC
Subject: நான்
Finite-verb: பார்த்தான்
Subject and finite-verb disagree in person (நான் is of first person but பார்த்தான் is of third person).
Suggestions: (பார்த்தேன்)
In this input sentence, the grammatical error detection system identifies the word 'நான்' as a subject and the word 'பார்த்தான்' as a finite-verb based on the main POS tags. The sub-tag FP of the subject indicates first person whereas the sub-tag TP of the finite-verb indicates third person causing the disagreement in person. In order to remove this disagreement, the correction system gives 'பார்த்தேன்' as a suggestion for the finite verb 'பார்த்தான்' based on the POS person sub-tags.

Example 4: Sentence with mistake in tense aspect agreement
Input sentence: அவன் நேற்று நண்பர்களுடன் பாடசாலைக்கு விரைவாக ஓடிப் போவான்.
POS tagger output: அவன்/PR-PRP-M-TP-SL நேற்று/N-NN-PA நண்பர்களுடன்/N-NN-PL-CM3 பாடசாலைக்கு/N-NN-SL-CM3 விரைவாக/RB ஓடிப்/V-VM-VNF-VBN போவான்/V-VM-VF-M-TP-SL-FU ./RD-PUNC
Subject: அவன்
Finite-verb: போவான்
Tense of the word 'நேற்று' disagrees with tense of the finite verb 'போவான்' ('நேற்று' indicates past tense but 'போவான்' indicates future tense).
Suggestions: (நேற்று/போவான்) or (நாளை/போவான்)
In this input sentence, the grammatical error detection system identifies the word 'அவன்' as a subject and the word 'போவான்' as a finite-verb based on the POS tags. The sub-tag PA in the tag N-NN-PA of the word நேற்று indicates past tense whereas the sub-tag FU in the tag V VM-VF-M-TP-SL-FU of the finite-verb indicates future tense causing the tense disagreement. In order to remove this tense disagreement, the correction system gives (நேற்று/போவான்) or (நாளை/போவான்) as suggestions based on the POS tense sub-tags.

This experiment is tested with two sets of sentences:

1. Set 1 consists of ten thousand sentences that are picked randomly from various sources online Tamil articles and Tamil news websites.
2. Set 2 consists of 150 sentences deliberately made with all kinds of grammatical mistakes: mistakes in subject and finite-verb agreement with respect to person, number and/or gender, and mistakes in tense aspects agreement.

The following table shows the details of

(A) the number of grammatical errors (mistakes in subject and finite-verb agreement with respect to person, number and/or gender, and mistakes in tense aspects agreement) existed in the chosen sets of sentences,
(B) the number of grammatical errors detected by the system,
(C) the number of grammatical errors with at least one suggestion, and
(D) the number of generated suggestions.

Test Set	(A)	(B)	(C)	(D)
Set 1	48	48	48	63
Set 2	150	150	150	206

In our testing with 10,000 sentences having 48 grammatical errors (mistakes in subject and finite-verb agreement with respect to person, number and/or gender, and mistakes in tense aspects agreement), the correction system gives 63 suggestions, a Scholar in Tamil language approved all suggestions as most suited. Moreover, for 150 sentences deliberately made with all kinds of grammatical mistakes mentioned above, it gives 206 suggestions. A Scholar in Tamil language approved 197 suggestions as most suited and the remaining 9 are acceptable.

3. CONCLUSION

In this work, performance of our grammar checker has been evaluated. In this regard, hierarchical POS tags of words are used to detect the grammatical mistakes and two sets of Tamil grammar rules are used to generate suggestions. Test results show that the grammar checker performs well in giving suggestions for mistakes detected as subject and finite-verb disagreement or mistakes in tense aspects agreement in Tamil sentences.

REFERENCES

[1] A. Navalar, Tamil Grammar Questions and Answers, No. 366, Kankesanthurai Road, Jaffna: Vannai Santhayarmadam, 1998.

[2] V. Sangar, Tamil Grammar, Puduchcheri, India: Nanmozi Printers, 2006.

[3] M. A. Nuhman, Basic Tamil Grammar, Kalmunai: Department of Tamil, University of Peradeniya: Readers Association, 2013.

[4] Kanesaiyar, Tholkappiyam, Chunnakam: Thirumagal Printers, 1937.

[5] A. Navalar, Nannool, 1994.

[6] K. Kukich, Techniques for Automatically Correcting Words in Text. ACM Computing Survey, vol. 24, no. 4, pp. 377--439, 1992.

[7] M. S. Gill, G. S. Lehal and S. S. Joshi, "A Punjabi Grammar Checker," in International Joint Conference on Natural Language Processing, 2008.

[8] Sobha, Sindhuja, Gracy, Padmapriya, Gnanapriya, and Parimala. AUKBC Tamil Part-of-Speech Corpus (AUKBC-TamilPOSCorpus2016v1). Computational Linguistics Research Group, AU-KBC Research Centre, Chennai, India, 2016.

[9] R. Sakuntharaj and S. Mahesan, A Refined POS Tag Sequence Finder for Tamil Sentences. in 2018 IEEE International Conference on Information and Automation for Sustainability (ICIAfS), Colombo, Sri Lanka, 2018.