Rule based Simple English Sentence Correction by Rearrangement of Words

Namrata Pratap Simha, Vishwas Manohar, Sudarshan Suresh M., Dheeraj D. Bhat, Dr. Saritha Chakrasali

Abstract— Natural language processing (NLP) is a field in computer science research that is exploring automation of spoken languages. This domain has a lot of potential to produce applications that will reduce ambiguity between humans and machines. Correction of English sentences given an incorrect sentence with words in the wrong order is one such application that enables people to learn English at a basic level through translation. In this work, a rule-based approach is employed to rearrange words in a wrongly ordered simple English sentence to obtain a correct and meaningful sentence. This is a prerequisite for any translation software.

Index Terms—Rule based training, Natural language processing, Part-of-speech tagging, English grammar, Word-types, Computational linguistics, Sentence autocorrect, Named-entity recognition.

1 Introduction

Word-type of a given word refers to the category it belongs to as defined by English grammar. This is used as a parameter to arrange sentences grammatically. The word types considered here are a few and their arrangement according to proper positions is proposed in this work. This will enable correction of simple sentences which contain word types that are considered in this work.

A major application of NLP is the translation of languages from one to another. In the machine translation of any other language to English, an interface is first used to input a sentence. The second step is to pre-process the input text and then parse it into corresponding English characters and then to corresponding English words. Subsequently, an algorithm is applied to get the final output in the form of correct English sentences [1]. During this process, after pre-processing the input text, incorrectly ordered English sentences are obtained. Correction of these sentences is a major issue to tackle.

Sentence correction has been an important emerging issue in computer-assisted language learning. Existing techniques based on grammar rules or statistical machine translation are still not robust enough to tackle the common errors in sentences produced by language learners. A relative position language model and a parse template language model has been proposed to complement traditional language modeling techniques in addressing this problem [2]. Also, prepositional phase errors and orthographic errors are the major issue in machine translation [3].

The correction of a sentence starts with tagging word-types to determine the word's position in the sentence. The tagging process is carried out through "part-of-speech tagging" techniques [4], [5]. In this work, simple word and word-type associations are used. Further, to tackle the issue of sentence correction, a rule based approach using the tagged word-types for the reordering of words is described in this work.

2 Procedure for Rule based Simple English Sentence Correction

In this work, a rule based approach is employed to handle the correction of English sentences by rearrangement. Using various grammar rules that tackle simple sentences from the basic grammar rules in [6], a rule has been devised to handle jumbled words in incorrect English sentences. These sentences can be rearranged after word-tagging to form simple meaningful English sentences.

In this work, each word is associated with a specific word-type (subject, object, etc.). Each one of these word-types is assigned a particular position in a sentence based on the rule devised. If the words belonging to particular word-types take positions other than the ones mentioned in the grammar rule, then the words are swapped to make the sentence grammatically correct, thereby essentially rearranging all the words to obtain a meaningful sentence.

3 Equations

The grammar rule to correct scrambled English sentences is given by the equation below:

$$\textit{(Subject adjective)–subject–verb–(object adjective)-} \\ \textit{(object)–[indeclinables list]–[(adjective)(noun) pairs list]} \quad (1)$$

The terms used in the equation are:

1. Subject adjective: The adjective for subject.
2. Subject verb: Verb that refers to the subject.
3. Object adjective: The adjective for object.
4. Object: Refers to the object in the sentence.
5. Indeclinable list: The list of indeclinables (mostly used as prepositions).
6. Adjective noun pairs list: List of adjective and noun pairs that appear in a sentence.

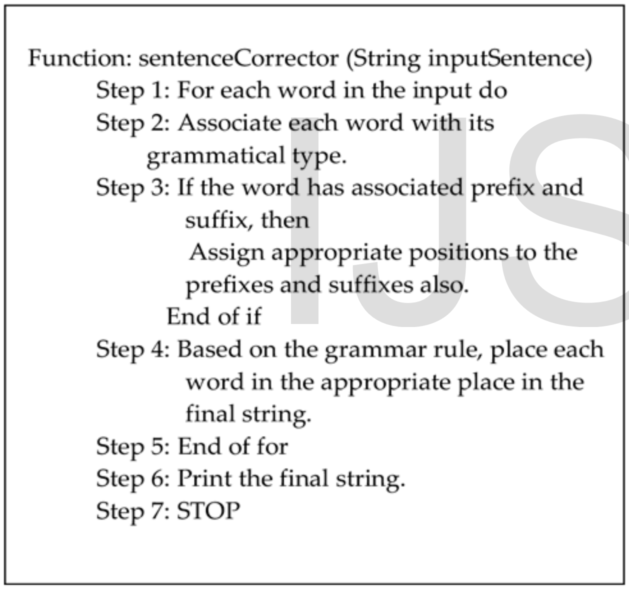
- *Namrata Pratap Simha is currently pursuing bachelor's degree program in Information Science and Engineering in BNM Institute of Technology, Bangalore, India, E-mail: namrata.simha@gmail.com*
- *Vishwas Manohar is currently pursuing bachelor's degree program in Information Science and Engineering in BNM Institute of Technology, Bangalore, India, E-mail: man.vishwas@gmail.com*
- *Sudarshan Suresh M. is currently pursuing bachelor's degree program in Information Science and Engineering in BNM Institute of Technology, Bangalore, India, E-mail: sudarshan1994@gmail.com*
- *Dheeraj D. Bhat is currently pursuing bachelor's degree program in Information Science and Engineering in BNM Institute of Technology, Bangalore, India, E-mail: bhat.dheeraj@gmail.com*
- *Dr. Saritha Chakrasali is currently working as a professor in the department of Information Science and Engineering, BNM Institute of Technology, Bangalore, India, E-mail: sarithachakrasali@bnmit.in*

4 Implementation and Results

During the implementation of the rule based sentence corrector, grammatically incorrect and unordered sentences were considered. Let us take for instance, the incorrect sentence “a pretty is girl Samantha.”. Such sentences have to be corrected so as to obtain a sensible and grammatically correct sentences.

Table 1 shows the pseudocode for the sentenceCorrector method. The sentence correction process is done using the ‘type of word’. Each word in the English dictionary is assigned a specific word-type, such as verb, noun, adjective, etc., which is usually already available. During autocorrect, the word type of each word is found. For instance, in the given example – “a pretty is girl Samantha” the word types would be: ‘a pretty’-adjective, ‘is’-verb, ‘girl’-noun, ‘Samantha’-proper noun. At this point, the adjective-noun pairs are determined based on their occurrence in the native language before translation to English. Thus, ‘a pretty’ and ‘girl’ become an adjective-noun pair.

Table 1: Pseudocode for sentenceCorrector function



```
Function: sentenceCorrector (String inputSentence)
    Step 1: For each word in the input do
    Step 2: Associate each word with its
            grammatical type.
    Step 3: If the word has associated prefix and
              suffix, then
              Assign appropriate positions to the
              prefixes and suffixes also.
           End of if
    Step 4: Based on the grammar rule, place each
              word in the appropriate place in the
              final string.
    Step 5: End of for
    Step 6: Print the final string.
    Step 7: STOP
```

There are basically 7 sets of words recognized in this work. They are: Subject, Subject adjective(s), Verb, Object, Object adjective(s), Indeclinable(s), Remaining Adjective - Nouns pairs. The assigned word-types of every word encountered in the input are compared with the 7 basic types. Proper nouns encountered are given preference to assign Subject or Object positions. So in the given example, the new assignment is made as follows: ‘a pretty’-object adjective, ‘is’-verb, ‘girl’-object, ‘Samantha’-subject. Every sentence will have words which come under at least a few of these word types, out of which subject and verb mandatorily appear, as shown in the example. Each of these word types has a specific position in an English sentence, so as to make the sentence grammatically correct. The order of word-types in the rule followed in this work is as follows:

1. Subject adjective(s)
2. Subject
3. Verb
4. Object adjective(s)
5. Object
6. Indeclinable(s)
7. Remaining Adjective - Nouns pairs.

This is as shown in (1). Now, once the compared word-types are matched, the words are placed in their respective positions in the sentence order. In the given example, the sentence placement is done as follows:

1. *null*
2. Samantha
3. is
4. a pretty
5. girl
6. *null*
7. *null*

The corrected sentence is finally obtained. The given example now becomes “Samantha is a pretty girl.”, which is a correct and meaningful English sentence.

5 Conclusion

This work takes into account the design and implementation of a system for the correction of jumbled sentences after translation of simple sentences from other languages to English. These sentences may consist of subject(s), object(s), verb, adjectives(s), noun(s) and indeclinable(s) using simple grammar rules, thus solving the issue of sentence correction that computer assisted language software faces today.

This work would help people by giving grammatically correct English sentences. This solution can be extended for different parts of a sentence with more complex word-tagging and grammar rules.

6 Acknowledgment

We would like to thank BNMIT for the facilities, guidance and encouragement. We would also like to thank our parents for all their support and encouragement.

7 References

[1] Pankaj Upadhyay, Umesh Chandra Jaiswal, Kumar Ashish “TranSish: Translator from Sanskrit to English-A Rule based Machine Translation” in the *International Journal of Current Engineering and Technology*, Vol. 4, No. 5 Oct. 2014.

[2] Chung-Hsien Wu, Chao-Hong Liu, Matthew Harris, Liang-Chih Yu “Sentence Correction Incorporating Relative Position and Parse Template Language Models” in the *IEEE Transactions on Audio, Speech, and Language Processing Journal*, Vol 18, Issue: 6.

[3] S. Suganthi, P. Bamarukmani, K. G. Srinivasagan, M. Saravanan “Semantic based orthographic with prepositional phrase for English-Tamil translation” in the *4th International Conference on Intelligent Human Computer Interaction (IHCI)*, Dec. 2012.

[4] Sharon Goldwater, Thomas L. Griffiths “A Fully Bayesian Approach to Unsupervised Part-of-Speech Tagging∗” in

the *Proceedings of ACL*, pages 744–751.

[5] Dan Garrette, Jason Baldridge “Learning a Part-of-Speech Tagger from Two Hours of Annotation” in the *Proceedings of NAACL*, Atlanta, Georgia.

[6] C. Wren and H. Martin. “High School English Grammar and Composition”.