I J C T A, 9(11) 2016, pp. 5119-5128
© International Science Press

A NOVEL APPROACH TO PARAPHRASE ENGLISH SENTENCES USING NATURAL LANGUAGE PROCESSING

Prateek Agrawal[*], Vishu Madaan[**], Nandini Sethi[***], Vikas Kumar[****] and Sanjay Kumar Singh[*****]

Abstract: This paper describes a novel approach to express any English sentence syntactically in different manner without changing its meaning. Our system mainly deals with sentences written either in present indefinite tense or past indefinite tense. It takes a sentence as input and produces another sentence without changing its semantic after applying Active-Passive voice conversion rules and synonyms /antonyms replacement method. Para-phrasing of sentences can be used to change a complex sentence in simplified form. In our paper, we have described complete algorithm to paraphrase the sentences. The performance of the system is totally dependent on the size of Database. This application can be helpful in designing robots to understand different forms of English sentences, to use as English tutor for students to get them idea about different form of sentences and in plagiarism tools to find the higher level of plagiarized text up to certain extent.

Key Words: Active passive voice conversion, synonyms, antonyms

1. INTRODUCTION

The term Paraphrasing is an importance application of natural language processing which means converting a sentence differently by keeping its semantic unchanged. In English language, it can be used to convert direct speech to indirect speech, active voice to passive voice and vice a versa. Here the actual meaning of the sentence is kept same as the original sentence. There is either change in the sequence of the word or, word's synonym or antonym is used. The reframing do not accompany the direct reference, it serves as a source to new reframed paragraph.

The scope of paraphrasing sentences is very vast. Sentences can be used to write the reviews on already available articles in simpler manner. It is also helpful in designing an intelligent system that can take decisions like humans.

Different ways to paraphrasing a sentence:

1. Conversion from active to Passive voice and vice a versa.
2. Conversion from direct speech to indirect speech and
3. vice a versa.

[*] prateek061186@gmail.com

[**] School of Computer Science Engineering, Lovely Professional University, Punjab (India) vishumadaan123@gmail.com

[***] nandinisethi2104@gmail.com

[****] vikas.cpp@gmail.com

[*****] sanjayksingh.012@gmail.com

4. Replacing words with its synonyms.
5. Replacing words with equivalent antonyms.
6. Rearranging the order of sentences on its priority etc

Various activities involved in the process of paraphrasing:

Text Segmentation

Firstly the text is divided into segments or different sentences. Sentence division is the issue of isolating a string of composed dialect into its segment sentences. In English and some different dialects, utilizing accentuation, especially the full stop character is a sensible estimate. However, even in English this issue is important because of the utilization of the full stop character for truncations, which might possibly additionally end a sentence. When preparing plain content, tables of truncations that contain periods can help forestall erroneous task of sentence limits.

Parts of Speech Tagging

Parsing or syntactic examination is the strategy of separating a progression of pictures, either in trademark tongue or in scripts, acclimating to the fundamentals of a formal sentence structure. The term has possibly differing ramifications in particular branches of semantics and programming building. Customary sentence parsing is routinely executed as a framework for understanding the exact significance of a sentence, on occasion with the backing of devices, for instance, sentence diagrams. It usually underlines the criticalness of semantic divisions, for instance, subject and predicate.

For example: John is a boy. He lives in Jalandhar. He is a student [14].

Sentence after POS tagging:

John_NNP is_VBZ a_DT boy_NN._. He_PRP lives_VBP in_IN Jalandhar_NNP._. He_PRP is_VBZ a_DT student_NN.

Mapping

Mapping is the drawing conclusions from the given premises to solve problems and make decisions. It basically manipulates the given knowledge and generates new knowledge from the defined rules and facts that are stored in the database. It derives the new knowledge with the help of logics or by using inference. So, the relationships that are stored in database are and will be used [11].

Reframing Rules

In the database the knowledge is represented in the forms of rules and facts. These facts and rules are applied to get the desired output. The rules can be certain patterns or symbols that matches with the input and for applying the processing.

Natural language processing (NLP) is a field of computer science, and artificial intelligence and scientific study of language that deals with the interactions between computers and human languages. There are many challenges for interaction between computers and humans. Computers understand only binary digits but humans can’t deal with binary digits. So we require huge database stored in our system for processing of human understandable words by computers. Natural Language Processing (NLP) is one of the technique through which humans can interact with computers [8].

1.1 STEPS OF NATURAL LANGUAGE PROCESSING

The process of NLP is performed by 5 Steps [6][7]:

1. **Morphological and Lexical Analysis:** It is the analysis of the structure and formation of the words and sentences.

7. **Syntax Analysis:** It is the analysis of the grammatical structure of the sentences. It is the study of sequence of words in the sentence as shown in Figure 1.

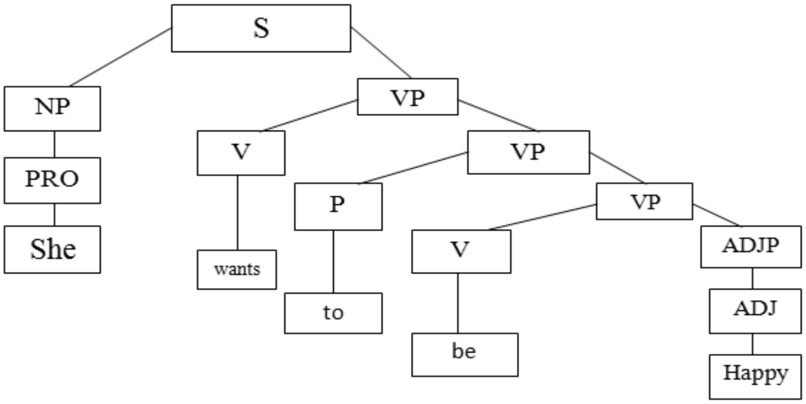


Figure 1: The result of syntactic analysis of "She wants to be happy"

8. **Semantic Analysis:** This step deals with the meaning of the sentences. It is the mapping between syntax and the meaning of the sentences. Semantic analysis is used to check whether inserted sentence is accurate or not. Although the main intend of semantic analysis is the formation of the target language representation of the sentence's meaning which indicate assigning meanings to the structures created by syntactic analysis.

9. **Discourse Integration:** It is the study of the dependency of one sentence on another or preceding sentence of the paragraph.

10. **Pragmatic Analysis:** This step involves the study of purposeful use of language or the world knowledge in particular situations.

1.2 Sentences

A sentence can incorporate words assembled definitively to express a statement, question, exclamation, demand, summon or suggestion. A sentence consists of words that on a basic level tell a complete thought, despite the fact that it may make little sense taken in disconnection outside of any relevant connection to the subject [12].

A sentence may consist of following components [10]:

1. Subject
2. Predicate
3. Clause
4. Phrase
5. Modifier

1. ***Subject:*** The subject of a sentence refers to an individual, thing, or thought which is doing or being something. The subject of a sentence is either a noun or a noun phrase.

   For example: "Ram is running."

   Here, "Ram" is the subject, because he is the actor in the sentence, who is doing something.

2. ***Predicate:*** There are mainly two parts of a sentence: one is subject and other is predicate. A predicate modifies the subject. The predicate consists of a verb, and the verb bounds that allow a sentence to complete in meaningful way.

   For example: "Ram is running."

   Here, "is running", is the predicate because it contains a verb and it is providing the information about the subject.

3. ***Clause:*** A clause tells some additional information about the subject.

   For example: "Ram is running fast."

   Here, "fast" is clause because it is adding additional information to subject.

4. ***Phrase:*** A phrase is like a dependent clause. A phrase is a group of words which can't be alone as a sentence, but it can add some information to a sentence.

   For example: "In the river"

   Here, "In the river" is a phrase and it can be attached to a sentence.

5. ***Modifiers:*** A modifier is a part of the sentence which adds certain information to the sentence or changes the actual meaning of a sentence.

2 . PREVIOUS WORKS

In the existing paraphrasing system such as Ginger software [9], developers have used a method using synonym, idioms and phrases replacement. There is no scenario for antonym and active to passive voice conversion. Marcel Bollmann (2014) performed syntax analysis in German language and for its implementation they focures on order of the words [3]. Lin and Pantel (2001) use a standard (nonparallel) monolingual corpus to generate paraphrases, based on dependency graphs and distributional similarity [4]. Pang et al. (2003) use parse trees over sentences in monolingual parallel corpus to identify paraphrases by grouping similar syntactic constituents [5]. Also most of the existing systems are static and only one form of new paragraph is given as output while in the proposed system there are different output screens for antonyms, synonyms, active to passive voice and one overall result including the these three things.

3. OUR APPROACH

Proposed system is done in various steps like text segmentation, tokenization, parts of speech tagging, classification of sentences, database generation, reframing rules etc shown in Figure 2:

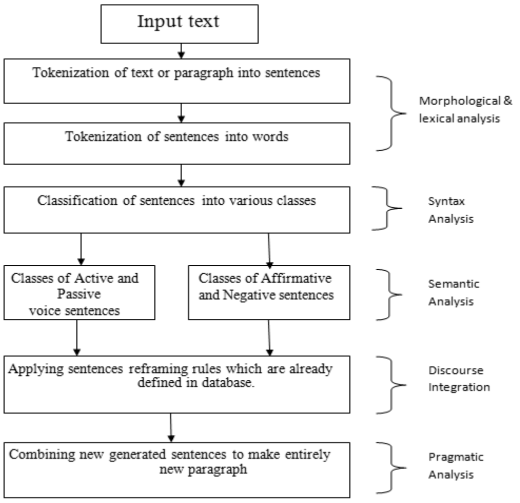

Input text
Tokenization of text or paragraph into sentences
Tokenization of sentences into words
Morphological & lexical analysis
Classification of sentences into various classes
Syntax Analysis
Classes of Active and Passive voice sentences
Classes of Affirmative and Negative sentences
Semantic Analysis
Applying sentences reframing rules which are already defined in database.
Discourse Integration
Combining new generated sentences to make entirely new paragraph
Pragmatic Analysis

Figure 2: Processing steps of proposed approach

3.1 Tokenization of text or paragraph into sentences:

Firstly the text is divided into segments or different sentences. Sentence division is the issue of isolating a string of composed dialect into its segment sentences. In English and some different dialects, utilizing accentuation, especially the full stop character is a sensible estimate. However, even in English this issue is important because of the utilization of the full stop character for truncations, which might possibly additionally end a sentence. When preparing plain content, tables of truncations that contain periods can help forestall erroneous task of sentence limits [2].

3.2 Tokenization of sentences into words:

Various grammatical rules are applied after text segmentation and POS tagging of sentences. In this system Stanford POS tagger is used. In this proposed system english-left3words-distsim tagger model is used to tag the words [14].

3.3 Classification of sentences into various classes:

Here, the sentences are classified into various classes such as active and passive sentences, affirmative and negative sentences etc. This classification will be helpful in paraphrasing of sentences which do not change the meaning of the sentence but change the wording of the sentence. This will also helpful in reframing a complex sentence into simpler one [1].

Sentences are divided into two classes:

1. ***Classes of Active and Passive voice sentences:*** General Rules for active to passive voice conversion are described in Table 1:

Table 1
Active and passive voice condition table

Active voice	*Passive voice*
The starting of the sentence is subject.	Subject is not used.
Verb form is used as per tense.	Always 3rd form of verb is used.
Subjective form of subject is used in starting.	Objective form of subject is used in end.
Helping verb is used according to tense and condition.	Helping verb is used according to tense and condition.

2. ***Classes of Affirmative and Negative sentences:*** One class include the sentences which are positive in their meaning i.e. which do not include any negative word. Matching case examples are

   (a) Mohan plays cricket.

   (b) Mohan played cricket.

   Another class include the sentences which are negative in their meaning i.e. which include any negative word.

   Matching case examples:

   (a) Mohan do not play cricket.

   (b) Mohan did not played cricket.

3.4 Applying reframing rules:

In the database the knowledge is represented in the forms of rules and facts. These facts and rules are applied to get the desired output. The rules can be certain patterns or symbols that matches with the input and for applying the processing. After input of a paragraph is segmented into various lines and POS tagging of each sentence is found. Then these tagging are matched with the specified rules which are already defined in the database. In this proposed system regex pattern matcher is used to match the pattern of a sentence and apply the rules accordingly.

One of the patterns defined in this system on which pattern matching is done to apply desired case is below:

Pattern for Simple Present and Past Tense

(|JJ)(|NNS|NNPS |NNP |NN)(NNS|NNPS|NNP|NN|PRP|PRP$) (VBZ|VBP|VB|VBD)
(.| DT | JJ)(NNS|NNPS|NNP|NN).

Where, JJ= Adjective

NNP= Proper Noun

PRP= Pronoun

VB= Verb

Matching case examples

Mohan plays cricket.

Mohan played cricket.

Synonyms Replacement

Two or more interrelated words that can be changed in a context are synonyms. In this system user will be provided with a list of options of synonyms to choose from. A user can select any of synonyms from given drop down list. The synonyms in the available list would be arranged according to similarity index. The word which is nearest to the matched words is replaced accordingly. Some of the examples of synonyms are illustrated in Table 2:

Table 2: Synonym Replacement

Word	*Sample synonyms*
Good	Beneficial , Goodness, Thoroughly, Well
Heaven	Eden, Paradise , Nirvana
World	Earth , Universe, Reality
Human	Human being , Man
Dull	Dumb , Muffled

Antonym Replacement

When a word expresses the meaning opposite to the given word, it is referred as antonym. In this proposed system negative sentences can be converted to positive sentences by using specific antonym. Some of the examples of antonyms are illustrated in Table 3:

Table 3: Antonym Replacement

Word	*Sample Replacement*
Not good	Bad
Not ugly	Beautiful , Lovely
Not poor	Rich
Not genuine	Counterfeit
Not bad	Good

Active – Passive voice Conversion

When the subject actively participates in sentences then that sentence is in active voice. These sentences can be converted to passive voice using various grammar rules. In active voice the subject always precedes the object. But in passive voice the object is before the subject. Active voice sentences is used when something is directly spoken, whereas passive voice is used when somebody else refer to that sentence. Some of the examples of Active-Passive Voice are illustrated in Table 4:

Table 4: Active Passive conversion

Tense	*Sentence*	*Voice*
Simple Present	Active	Mohan plays cricket.
Simple Present	Passive	Cricket is played by Mohan.
Simple Past	Active	Mohan ate a mango.
Simple Past	Passive	A mango was eaten by Mohan
Present Continuous	Active	John is watching a movie.
Present Continuous	Passive	A movie is being watched by John.

3.5 Combining new generated sentences to make entirely new paragraph

The reframing is the restatement of the paragraph without changing the meaning of the actual text. It explains the paragraph in simple words and enables to present the same paragraph in different ways.

Following Figure 3 shows the example that explains the meaning of reframing:

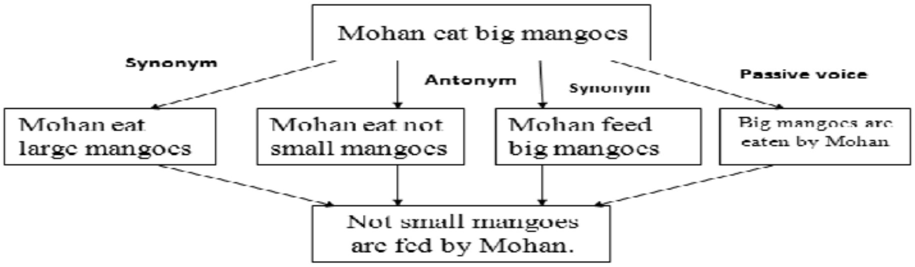


Figure 3: Reframing types

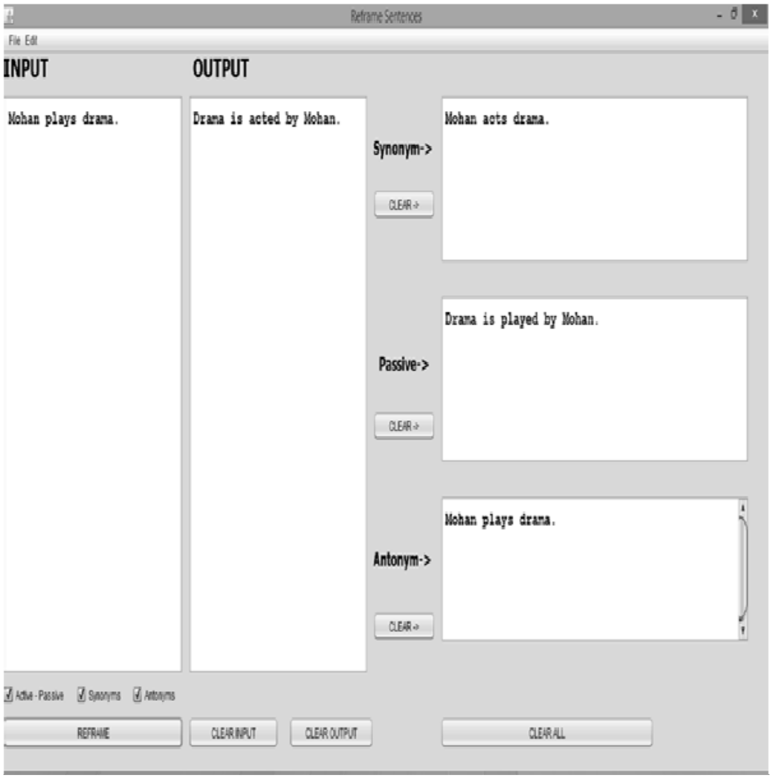
Here, the sentence “Mohan eats big mangoes” is, reframed using synonym replacement rules, antonym replacement rules and active to passive voice conversion rule. Here the actual meaning of the sentence is kept same as the original sentence. There is either change in the sequence of the word or, word's synonym or synonym is used. The reframing do not accompany the direct reference, it serves as a source to new reframed paragraph.

Proposed algorithm

```
Paraphrase (sentence or paragraph)
INPUT Text directly or through text file.
Apply Co reference resolution by replacing pronoun with its individual entity.
Apply Text Segmentation.
Apply POS tagging to individual sentences
Set count=1
While (count < number_of_sentences)
{
Convert Active to passive voice as per matched pattern from database
Replace similar words with Synonym.
Replace negative words with Antonyms
}
Count++
Return Output
End
```

4. RESULTS

This work will generate new sentences based on specified rules and after synonyms and antonyms replacements. Figure 4 shows the main interface of our proposed system and results are illustrated in Table 5.



INPUT: Mohan plays drama.

OUTPUT: Drama is acted by Mohan.

Synonym-> Mohan acts drama.

Passive-> Drama is played by Mohan.

Antonym-> Mohan plays drama.

☑ Active - Passive ☑ Synonyms ☑ Antonyms

Figure 4: Interface of the System

Table 5
Result Analysis

Input	*Replacement and Reframing*	*Result*
The world is beautiful.	Synonym	The world exists beautiful.
	Antonym	The world is not ugly
	Synonym +Antonym	The world exists not ugly.
Mohan plays drama.	Synonym	Mohan acts drama.
	Active to passive	Drama is played by Mohan.
	Synonym +Active Passive	Drama is acted by Mohan.
I do not like sour mangoes.	Antonym	I like sweet mangoes.
	Active to passive	Sour mangoes are not liked by me.
	Antonym +Active Passive	Sweet mangoes are liked by me.
John does not eat small mangoes.	Synonym	John does not feed little mangoes.
	Antonym	John eats large mangoes.
	Active Passive	Small mangoes are not eaten by John.
	Synonym +Antonym +Active-Passive	Large mangoes are eaten by John.

5 CONCLUSION

In this proposed system we have discussed sentence reframing technique using NLP. This system can be used by scholars, technical writers and researchers. This system can be further extended to develop software for semantic analysis for information extraction and other. This system can be helpful in making a robot understand different forms of sentences. It can also be used to develop an intelligent system that can take decisions like humans. This work can be extended to make a decision support system that will work is a similar manner like humans and can respond just like humans by understanding the different forms of sentences given to it as an input.

References

[1] Pera, Maria Soledad; and Ng, Yiu-Kai; "Classifying sentence based summaries of web documents", 21st IEEE International Conference on Tools with Artificial Intelligence, 2009, pp. 443-440

[2] Minnen, Guido; Carroll, John; Pearce, Darren; "Applied morphological processing of English", Natural Language Engineering, vol. 7, issue 03 (2001), pp. 1-18.

[3] Bollman, Marcel; "Adapting SimpleNLG to German", Proceedings of the 13th European Workshop on Natural Language Generation (ENLG-11), (2011), pp. 133-138

[4] Lin, Dekang; Pantel, Patrick; "DIRT – discovery of inference rules from text", In *Proceedings of ACM SIGKDD Conference on Knowledge Discovery and Data Mining,* 2001, pp. 323-328

[5] Pang, Bo; Knight, Kevin; Marcu, Daniel; "Syntax-based alignment of multiple translations: Extracting paraphrases and generating new sentences", In *Proceedings of HLT/NAACL*, 2003, pp. 102-109

[6] D.W.Patterson, Introduction to AI & Expert Systems, Prentice Hall.

[7] Rich, Knight," Artificial Intelligence", Tata McGraw Hill, 2009(Third edition).

[8] Steven Bird, Ewan Klein, and Edward Loper (2009 , July) Natural Language Processing with Python, O"Reilly Media , 1st edition.

[9] Ginger-Software // accessed on 15th January 2016 http://www.gingersoftware.com/products/sentence-rephraser#. VroNCfl97IU

[10] Parts-of-Sentence // accessed on 16th January 2016 http://www.really-learn-english.com/parts-of-a-sentence.html

[11] NLP. //accessed on 10th Jan 2016 http://www.mind.ilstu.edu/curriculum/protothinker/natural_language_processing.php

[12] Sentence definition, // accessed on 2nd February 2016 https://www.englishclub.com/grammar/what-is-a- sentence.htm

[13] Subject-Verb-Agreement // accessed on 7th January 2016 http://grammar.ccc.commnet.edu/grammar/sv_agr.htm

[14] Standford POS tagger // accessed on 4th February 2016 http://nlp.stanford.edu/wiki/Main_Page