ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 8, Issue 5, May 2019

# Text Simplification Based On Lexical Analysis

Aysha Shahistha[1], Gangothri P Nair[2], Kavya M[3], Sherin K S[4], Anjali S[5]

U.G. Student, Department of Computer Engineering, Model Engineering College, Thrikkakara, Cochin, India[1,2,3,4]

Assistant Professor, Department of Computer Engineering, Model Engineering College, Thrikkakara, Cochin, India[5]

**ABSTRACT**: Text Simplification is the process of modifying natural language to reduce its complexity and improve both readability and understandability. It may involve modifications to the syntax, the lexicon or both. Text simplification is within the field of natural language processing is very similar to other techniques such as machine translation, text summarisation and paraphrase generation. Text simplification is different to text summarization as the focus of text summarisation is to reduce the length and content of input. While simplified texts are typically shorter, this is not necessarily the case and simplification may result in longer output, especially when generating explanations. Text simplification consist of two components: Syntactic simplification and Lexical simplification. While syntactic simplification aims at reducing the grammatical complexity of a sentence, lexical simplification focuses on replacing difficult words or short phrases by simpler variants.

**KEYWORDS**: Text simplification, readability, understandability.

## I. INTRODUCTION

Text is a fundamental part of our daily interaction with the information world. If text is simplified for an end user, then this may improve their experience and quality of life. Text Simplification is an NLP task that aims to rewrite sentences to reduce their syntactic or lexical complexity while preserving their meaning. Automatic text simplification is a research field in computational linguistics that studies methods and techniques to simplify textual content. Text simplification methods should facilitate or at least speed up the adaptation of available and future textual material, making accessible information for all a reality. Readability and understandability are related and a text which is easier to read is likely to be more understandable, as the reader will find it easier to take the time to look over the difficult concepts.

Paper is organized as follows. Section II describes related works. Algorithms used in different stages of text simplification is given in Section III. Section IV presents conclusion. Finally, Section V presents reference.

## II. RELATED WORK

Papers related to text simplification are as follows:

### A. An architecture for a text simplification system

A pipelined architecture for a text simplification system is presented and the implementation of the three stages analysis, transformation and regeneration is described. The analysis stage performs two functions. Firstly, it evidence of the relation between word lengths and lexical identifies syntactic structures that can be simplified. This can be done reliably using pattern matching techniques on POS tagged text with noun and verb groups. Once a sentence has been identified as suitable for simplification, the analysis stage provides the transformation stage with the required analysis. The first stage provides a structural representation of a sentence and the second stage uses a sequence of rules to transform this representation and attend the resulting structures into plain text. Particular emphasis is laid on the discourse level aspects of syntactic simplification as these are crucial to the process and have

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 8, Issue 5, May 2019

not been dealt with by previous research in the field. These aspects include generating referring expressions, deciding determiners, deciding sentence order and preserving rhetorical and anaphoric structure.

B. **A Survey of Automated Text Simplification**

A survey on automated text simplification is performed. There are many approaches to the simplification task, including: lexical, syntactic, statistical machine translation and hybrid techniques. This survey also explores the current challenges which this field faces. Lexical simplification is the task of identifying and replacing complex words with simpler substitutes. Syntactic simplification is the technique of identifying grammatical complexities in a text and rewriting these into simpler structures. Explanation generation is the technique of taking a difficult concept in a text and augmenting it with extra information, which puts it into context and improves user understanding. Statistical machine translation involves automatic techniques to convert the lexicon and syntax of one language to that of another, resulting in translated text. Text simplification is not solely confined to the reader, it may also be applied by the author to a text in order to ensure his point is clearly communicated, or even in a natural language processing pipeline to improve the performance of later components.

C. **Complex Word Identification: Challenges in Data Annotation and System Performance**

The problem of complex word identification (CWI) following up the SemEval CWI shared task has been considered and analysed. The use of ensemble classifiers to investigate how well computational methods can discriminate between complex and non-complex words made the task a little bit easier. At First, evaluate the dataset annotation and the performance of systems participating in the SemEval CWI task. It estimates the theoretical upper bound performance of the task given the output of the SemEval systems. Secondly, investigate whether human annotation correlates to the systems performance by carefully analysing the samples of multiple annotators. The dataset compiled for the shared task contains a training set composed of 2,237 instances and a test set of 88,221 instances. The first goal is to build high performance classifiers using plurality voting and second is to estimate the theoretical upper bound performance given the output of the systems that participated in the SemEval CWI competition using the oracle classifier. Complex words in the dataset are on average 6.74 characters long whereas non-complex words are on average 5.94 characters long. Experimental results provided empirical complexity for this dataset.

D. **Motivations and methods for text simplification**

To simplify a sentence, we need an idea of the structure of the sentence, to identify the components to be separated out. Simplification is a two stage process. The first stage provides a structural representation for a sentence on which the second stage applies a sequence of rules to identify and extract the components that can be simplified. A parser could be used to obtain the complete structure of the sentence. However, full parsing is slow and prone to failure, especially on complex sentences. There are two alternatives to full parsing which could be used for simplification. The first approach uses a Finite State Grammar (FSG) to produce noun and verb groups while the second uses a super tagging model to produce dependency linkages. These parsers are fast and reasonably robust; they produce sequences of noun and verb groups without any hierarchical structure. Compare the two approaches, and address some general concerns for the simplification task.

E. **An overview on text coherence methods**

An exhaustive survey which intends to investigate some of the most relevant approaches both in the areas of semantic and syntactic text coherence recognition methods, giving special emphasis to empirical methods and syntactic techniques. The increasing availability of texts generated in many area and online information has necessitated intensive research in the area of automatic text coherence identification within the Natural Language

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 8, Issue 5, May 2019

Processing (NLP) community. The problem has been addressed from many different perspectives, in varying domains and using various paradigms such as text summarization, text simplification and text generation.

F. **Text simplification using synchronous dependency grammars**

Describes a hybrid system that performs text simplification using synchronous dependency grammars.Main contributions in this approach are, a study of how automatically derived lexical simplification rules can be generalised to enable their application in new contexts without introducing errors, and an evaluation of our hybrid system that combines a large set of automatically acquired rules with a small set of handcrafted rules for common syntactic simplification. The evaluation shows significant improvements over the state of the art, withscores comparable to human simplifications. Present a unified framework for representingrules for syntactic and lexical simplification, and study how the definition of context affects system performance.

G. **Text Simplification Tools: Using Machine Learning to Discover Features that Identify Difficult Text**

Presented a simple and efficient splitting algorithm based on an automatic semantic parser. After splitting, the text is amenable for further simplification operations. The main structural simplification operation is rewriting a single sentence into multiple sentences while preserving its meaning. After splitting, NMT based simplification is performed, using the NTS system. The methods for performing sentence splitting as pre-processing allows the text simplification system to perform other structural and lexical operations, thus increasing both structural and lexical simplicity. Presented the first simplification system combining semantic structures and neural machine translation, showing that it outperforms existing lexical and structural systems. The consideration of sentence splitting as a decomposition of a sentence into its Scenes is further supported on structural TS evaluation.

H. **Readability Assessment for Text Simplification**

Readability assessment approach to support the process of text simplification for poor literacy readers. Given an input text, the goal is to predict its readability level, which corresponds to the literacy level that is expected from the target reader: rudimentary, basic or advanced. Readability assessment tool, developed as part of the PorSimples project, SIMPLIFICA offers simplification operations addressing a number of lexical and syntactic phenomena to make the text more readable. Readability assessment tool, the author is able to automatically check complexity/readability level of the original text. .Readability assessment is useful not only to inform authors preparing simplified material about the complexity of the current material, but also to guide automatic simplification systems to produce simplifications with the adequate level of complexity according to the target user.

I. **A Context-Aware Approach for the Identification of Complex Words in Natural Language Texts**

In order to map the semantics of natural language texts into a machine-readable format, it is important to understand when two textual elements (paragraphs, sentences, phrases or even words) have the same meaning. The effect of the context on the identification of complex words in natural language texts is evaluated. The approach automatically tags words as either complex or not, based on two sets of features: base features that only pertain to the target word, and contextual features that take the context of the target word into account. Experiments were done with several supervised machine learning models, and trained and tested the approach with the SemEval-2016 dataset. Results show that contextual features are just as important as features pertaining to the target word alone, and that considering them can significantly improve the recognition of complex words.

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 8, Issue 5, May 2019

## III. ALGORITHMS

**Lexical Analysis**

It identifies and replaces complex words with simpler substitutes. This involves no attempt to simplify the grammar of a text but instead focuses on simplifying complex aspects of vocabulary.

Input : Text to be simplified
Output: Simplified Text

- Identification of complex words
- Substitution generation
- Synonym ranking

**Complex word identification:**

Complex word is identified based on

1. **Length**
   Word length, measured in either characters or syllables is a good indicator of complexity. In informal speech and writing short words are usually chosen over longer words for efficiency. This means that we are in contact more often with shorter words than we are with longer words and may explain in part the correlation between length and complexity.

2. **Morphology**
   Longer words tend to be made up of many parts - referred to as morphological complexity. The more parts there are, the more complex the word will be.

3. **Familiarity**
   The frequency with which we see a word is thought to be a large factor in determining lexical complexity. We are less certain about the meaning of infrequent words, so greater cognitive load is required to assure ourselves we have correctly understood a word in it's context. Familiarity is typically quantified by looking at a word's frequency of occurrence in some large corpus.

4. **Ambiguity**
Certain words have a high degree of ambiguity. A reader must discern the correct interpretation from the context around a word. This can be measured empirically by looking at the number of dictionary definitions given for a word.

5. **Context**
Context also affects complexity. Words in familiar contexts are more simple than words in unfamiliar contexts. This indicates that a word's complexity is not a static notion, but is influenced by the words around it. This can be modelled, using n-gram frequencies to check how likely a word is to co-occur with those words around it.

**Substitution generation**:

The identified complex word is replaced with its synonym. Two datasets Wordnet and Word2vec are used to generate synonyms of the complex word identified.

Input : Text to be simplified and complex word
Output : Set of synonyms for the complex word

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

*(A High Impact Factor, Monthly, Peer Reviewed Journal)*

*Visit:* www.ijirset.com

Vol. 8, Issue 5, May 2019

1. word2vec_synonyms=wordvec_model.most_similar(complex_word)
2. wordnet_synonyms=[]
3.for syn in wordnet.synsets(complex_word):
4. for lemma in syn.lemmas():
5. wordnet_synonyms.add(lemma.name())
6. set_of_synonym=word2vec_synonyms + wordnet_synonyms

**Ranking of synonyms:**

A set of substitutes are obtained for the complex word. The set of synonyms are first filtered. Then the remaining synonyms are ranked based on context and frequency. Then the least complex word is chosen as the substitute for the complex word identified.

Input : set of synonyms and the sentence to be simplified
Output : best substitute for the complex word is identified and replaced

1. newwords=set_of_synonyms
2. best_substitutes = {word : newwords [word] for word in newwords if self.checkWordFitContext(sentence, complex_word, word) and if postagCheck(sentence,index,word)}
3. best= max_frequency_in(best_substitutes)
4. replace(complex_word, best)

## IV. CONCLUSION

Text simplification is one of the research specific area under natural language processing. Many more up gradations are still going on this topic. It can be classified into two groups, syntactic and lexical simplification. Lexical Simplification is studied in this paper. In this study, important design criteria of text simplification system are presented.

## REFERENCES

[1] Marcos Zampieri1, Shervin Malmasi, Gustavo Paetzold, Lucia Specia ”Complex Word Identification: Challenges in Data Annotation and System Performance”,Proceedings of the 4th Workshop on Natural Language Processing Techniques for Educational Applications, pp.59–63, 2017.
[2] A. Siddharthan, ”An architecture for a text simplification system,” Language EngineeringConference,pp. 64-71, 2002.
[3] Matthew Shardlow,”A Survey of Automated Text Simplification”,(IJACSA) InternationalJournal of Advanced Computer Science and Applications, Special Issue on Natural LanguageProcessing, pp.58-67, 2014.
[4] E. Davoodi, L. Kosseim and M. Mongrain, ”A Context-Aware Approach for the Identificationof Complex Words in Natural Language Texts,” IEEE 11th International Conference onSemantic Computing (ICSC), San Diego, pp. 97-100, 2017.
[5] Sandra Aluisio, Lucia Specia, Caroline Gasperin,and Carolina Scarton,”Readability Assessment for Text Simplification ”, Proceedings of the NAACL HLT Fifth Workshop on Innovative Use of NLP for Building Educational Applications, pp. 1–9, 2010.
[6] D. Kauchak, O. Mouradi, C. Pentoney and G. Leroy, “Text Simplification Tools: UsingMachine Learning to Discover Features that Identify Difficult Text,” 47th Hawaii International Conference on System Sciences, Waikoloa, pp. 2616-2625, 2014.
[7] Mohamad Abdolahi and Morteza Zahedi,”An overview on text coherence methods”,Eighth International Conference on Information and Knowledge Technology (IKT), Hamedan,Iran,pp.1-5 , 2016.
[8] M.A. Angrosh and Advaith Siddharthan, ”Text simplification using synchronous dependencygrammars: Generalising automatically harvested rules”, Proceedings of the 8th International Natural Language Generation Conference, pp.16–25, 2014.
[9] Elior Sulem, Omri Abend, Ari Rappoport,”Simple and Effective Text Simplification UsingSemantic and Neural Methods”, Department of Computer Science, The Hebrew University of Jerusalem, pp.1-9, 2017.
[10] R.Chandrasekar, Christine Doran, B. Srinivas, ”Motivations and methods for text simplification”, University ot7 Pcnnsylwmia, lqfiladclphia,pp.1041-1044, 2015.
[11] HoracioSaggion, Graeme Hirst, "Automatic Text Simplification", in Automatic Text Simplification, Morgan & Claypool, pp.7-31, 2017.
[12] Steven Bird, Ewan Klein, and Edward Loper, “Natural Language Processing with Python - Analyzing Text with the Natural Language Toolkit”, Published by O’Reilly Media, pp.79-334,2009.
[13] Nitin Hardeniya, Jacob Perkins, Deepti Chopra, Nisheeth Joshi, Iti Mathur, “Natural Language Processing: Python and NLTK”, Packt publishing limited, pp.195-212, 2016.