Developing a Monolingual Sentence Simplification Corpus for Urdu

Yusra Anees
Fatima Jinnah Women University / Pakistan
yusra.anees96@gmail.com

Sadaf Abdul Rauf
Fatima Jinnah Women University / Pakistan
CNRS-LIMSI / France
sadaf.abdulrauf@gmail.com

Nauman Iqbal
Capital University of Science and Technology / Pakistan
nauman@biit.edu.pk

Abdul Basit Siddiqi
Capital University of Science and Technology / Pakistan
abasit.siddiqui@cust.edu.pk

Abstract

Complex sentences are a hurdle in the learning process of language learners. Sentence simplification aims to convert a complex sentence into its simpler form such that it is easily comprehensible. To build such automated simplification systems, corpora of complex sentences and their simplified versions is the first step to understand sentence complexity and enable the development of automatic text simplification systems. No such corpus has yet been developed for Urdu and we fill this gap by developing one such corpus to help start readability and automatic sentence simplification research. We present a lexical and syntactically simplified Urdu simplification corpus and a detailed analysis of the various simplification operations. We further analyze our corpora using text readability measures and present a comparison of the original, lexical simplified, and syntactically simplified corpora.

1 Introduction

Research in the last decade has been focusing on identification of complexity levels of sentences so that complexity of such sentences can be reduced to facilitate learning for students as per their learning grade. This is specifically true for Urdu for which this gap is increasing day by day, literary texts often include complex words and composite sentence structure (Alison and Mushta,2004). Our focus will be on such language; Urdu. In fact, no such prior work or resource exists for Urdu. It is the need of the day to address this issue and come up with effective complexity reduction and readability enhancement measures.

To enable research on automatic text simplification systems and text readability for Urdu, development of a simplification corpus providing enough complex sentences and their corresponding simple versions is imperative. We have developed one such corpus for the high school students and simplified (lexical and syntactically) short stories from a renowned author. We have considered three-levels in our simplification process: Original, lexical simplified and syntactic simplified. In Lexical Simplification (LS), complex words are replaced with simple and easy words. Whereas, Syntactic Simplification (SS) may result in an entirely new but simpler sentence. Such sentence aligned texts have been prepared for many languages, for example PWKP (Zhu et al., 2010), Newsela (Xu et al., 2015), Onestop (Vajjala and Lucic, 2018) and SimPA (Scarton et al., 2018) for English. Sentence simplification corpora for other languages include Ancora (Taulé et al., 2008), ERNESTA (Barbu et al., 2015), CLEAR (Grabar and Cardon, 2018) etc.

Another contribution of our work is a detailed analysis of several readability metrics and their application to Urdu using our corpus. We computed the readability measures with the popular readability metrics FKGL, FRE, ARI and SMOG. For each of the corpora, i.e. original, lexical simplified and syntactically simplified lexical analysis has been done and the scores show correlation with human evaluations.

2 Corpus Development and annotation scheme

The data for current study has been gathered from the Urdu digital library [1]. It includes 69 short stories. The Target audience of this data is young to old age. The complex sentence structure with typical

[1] http://www.udb.gov.pk/

Urdu literature vocabulary has been used which was not easy to comprehend. Online Urdu Lughat [2] (dictionary) is used to find simpler synonyms.

Complex sentences has been processed for removal of irrelevant characters and words to avoid ambiguities in data-set. Simplified corpora is(are) rechecked by language experts to remove any anomalies. 4 Language Experts (Urdu native speakers) has manually annotated the corpus into the simplified form using two techniques: lexical and syntactic simplification. Simplified versions of the complex texts are produced by annotating each sentence, first lexical simplification then syntactic simplification (or the previous given) . There, the syntactic simplification included insertion, deletion, splitting, merging, and reordering are used to produce simpler sentences which are the most productive simplification operations according to the literature. The guideline has been given to the annotator is that in syntactic simplification is involves in the removal of phrases or words such that the main context and meaning of the sentence remains the same. It changes the order of words grammatically and inserts new words to reduce complexity. Merging and splitting of sentences are also used to reduce the complexity which is frequently used by (Zhu et al., 2010).

And, In lexical simplification, two operations are performed paraphrasing of difficult words or phrases with simple words or phrases. This operation is applied almost in all manually developed corpus mentioned in the instruction section to carry out simplified corpus. The syntactic simplification was applied at the top where, the lexical simplification had been processed. And 18.3% sentence are only syntactically simplified because these sentence had no lexical complexity. In 3 annotators labeled the data on the lexical and syntactic operations.

Our corpus creation methodology is consistent with the recent works like (Štajner et al., 2019; Scarton et al., 2018; Katsuta and Yamamoto, 2018; Grabar and Cardon, 2018; Brunato et al., 2016; Brunato et al., 2015) who also have simplified using basic lexical simplification operations and (Yatskar et al., 2010) for syntactic simplification. By Human Evaluation of the simplified sentences, annotators have ensured that the simplified sentences had the consistent level of simplification. For evaluation, Two Urdu annotators have annotated 10% of our corpus. The same Evaluation scheme was followed by (Sulem et al., 2018).

3 Simplification statistic

We have produced a corpus of 1220 simplified sentences by simplifying 610 sentences, both lexical and syntactical. After in-depth analysis of language and content, we have approximately 58.8% sentences which were lexical and syntactical simplified, 10% sentences were not very complex and only Lexical operation was sufficient to produce the final simplified version, whereas 18.3% sentences could only be simplified by Syntactic operations. Around 12.7% sentences were simple enough not to require simplification of any form as shown in Figure1.

Figure 2 shows that in our simplification scheme, rewording is the most significant operation through which 77.61% of lexical simplification was accomplished. Same trend was observed by (Coster and Kauchak, 2011) they were report 65% rewording operations for English. In case of Syntactic Simplification, deletion was found to be the most frequent operation accounting to 84% of cases, this is also in line with results from previous researches (Coster and Kauchak, 2011; Brunato et al., 2016; Gonzalez-Dios et al., 2018). Insertion, split and merge and reordering follow with 9.12%, 4.24% and 2.14% usage respectively.

4 Text Simplicity and Readability scales

Readability metrics are used to evaluate complexity of text by using mathematical formula. We chose Flesch Reading Ease and Flesch-Kincaid Grade Level (Flesch, 1948; Kincaid et al., 1975), SMOG McLaughlin, (Mc Laughlin, 1969) and Automated Readability Index (ARI) (Senter and Smith, 1967) to evaluate our corpus. Which are more generalized and can evaluate complexity on the basis of some basic parameters such as average sentence length, average word length and number of syllables in a word. Flesh Readability Ease (FRE) metric scores range between 0 and 100. Higher value means text is easy to read and lower value means higher the difficulty.

[2] http://www.urdulibrary.org/

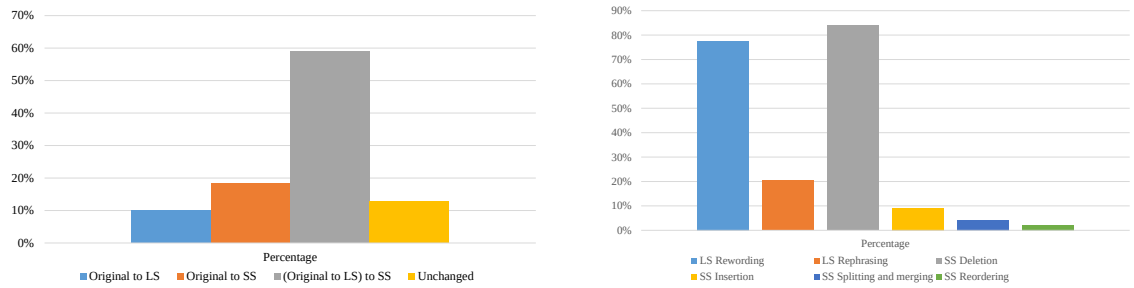

Figure 1: Percentages of sentences requiring different simplification procedures to get the final simplified sentences. The Figure 2: shows the percentage of each operation applied. LS indicates Lexical simplification and SS indication Syntactic simplification

5 Results Analysis

The metrics chosen for analysis, scores for FKGL, ARI and SMOG score are directly proportional to complexity, higher score means complex text and lower score means simpler text. However, for FRE this relation is inverse and a higher score means simpler text.

In Table 1 and Table 2 , We have some very interesting observations. The lowest score on FKGL is 6 for Syntactic simplification, Lexical simplification has score of 9, making it fall in the "average" complexity class and original complex sentence has a score of 12, which also puts it in the "average" complexity class but this is the threshold of the average complexity level for FKGL. We can claim that FKGL scores correctly identified the complexity level of text. As Table 2 shows that in case of FRE and ARI we see that even difference of small points is important in categorizing the level of text. SMOG scores of 4.12, 4.19 and 3.13 for complex, Lexical simplification and Syntactic simplification do not agree with the previous trend where and complex texts have 0.07 less score than LS, however Syntactic simplification has the lowest score. But this makes SMOG an unreliable metric for Urdu.

Metric	**Range**	**Level**
FRE	0 - 30	Skilled
	60 - 70	Average
	90 - 100	Basic
FKGL	13 - 18	Skilled
	7 - 12	Average
	1 - 6	Basic
ARI	13 - 18	Skilled
	7 - 12	Average
	1 - 6	Basic
SMOG	111 - 240	Skilled
	13 - 110	Average
	1 - 12	Basic

Table 1: Score range for readability metrics

	FKGL	**F**RE	**A**RI	**S**MOG
Original	12	91.78	6.39	4.12
Lexical	9	91.16	6.13	4.19
Syntactic	6	98.87	4.95	3.13

Table 2: Average scores of original and simplified sentences against FKGL, FRE, ARI and SMOG.

References

Eduard Barbu, M Teresa Martín-Valdivia, Eugenio Martínez-Cámara, and L Alfonso Ureña-López. 2015. Language technologies applied to document simplification for helping autistic people. *Expert Systems with Applications*, 42(12):5076–5086.

Dominique Brunato, Felice Dell'Orletta, Giulia Venturi, and Simonetta Montemagni. 2015. Design and annotation of the first italian corpus for text simplification. In *Proceedings of The 9th Linguistic Annotation Workshop*, pages 31–41.

Dominique Brunato, Andrea Cimino, Felice Dell'Orletta, and Giulia Venturi. 2016. Paccss-it: A parallel corpus of complex-simple sentences for automatic text simplification. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 351–361.

William Coster and David Kauchak. 2011. Simple english wikipedia: a new text simplification task. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: Short papers-Volume 2*, pages 665–669. Association for Computational Linguistics.

Rudolph Flesch. 1948. A new readability yardstick. *Journal of applied psychology*, 32(3):221.

Itziar Gonzalez-Dios, María Jesús Aranzabe, and Arantza Díaz de Ilarraza. 2018. The corpus of basque simplified texts (cbst). *Language Resources and Evaluation*, 52(1):217–247, Mar.

Natalia Grabar and Rémi Cardon. 2018. CLEAR – simple corpus for medical French. In *Proceedings of the 1st Workshop on Automatic Text Adaptation (ATA)*, pages 3–9, Tilburg, the Netherlands, November. Association for Computational Linguistics.

Akihiro Katsuta and Kazuhide Yamamoto. 2018. Crowdsourced corpus of sentence simplification with core vocabulary. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC-2018)*.

J Peter Kincaid, Robert P Fishburne Jr, Richard L Rogers, and Brad S Chissom. 1975. Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel.

G Harry Mc Laughlin. 1969. Smog grading-a new readability formula. *Journal of reading*, 12(8):639–646.

Carolina Scarton, Gustavo Paetzold, and Lucia Specia. 2018. Simpa: A sentence-level simplification corpus for the public administration domain. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC-2018)*.

RJ Senter and Edgar A Smith. 1967. Automated readability index. Technical report, CINCINNATI UNIV OH.

Sanja Štajner, Horacio Saggion, and Simone Paolo Ponzetto. 2019. Improving lexical coverage of text simplification systems for spanish. *Expert Systems with Applications*, 118:80–91.

Elior Sulem, Omri Abend, and Ari Rappoport. 2018. Semantic structural evaluation for text simplification. *arXiv preprint arXiv:1810.05022*.

Mariona Taulé, Maria Antònia Martí, and Marta Recasens. 2008. Ancora: Multilevel annotated corpora for catalan and spanish. In *Lrec*.

Sowmya Vajjala and Ivana Lucic. 2018. Onestopenglish corpus: A new corpus for automatic readability assessment and text simplification.

Wei Xu, Chris Callison-Burch, and Courtney Napoles. 2015. Problems in current text simplification research: New data can help. *Transactions of the Association of Computational Linguistics*, 3(1):283–297.

Mark Yatskar, Bo Pang, Cristian Danescu-Niculescu-Mizil, and Lillian Lee. 2010. For the sake of simplicity: Unsupervised extraction of lexical simplifications from wikipedia. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 365–368. Association for Computational Linguistics.

Zhemin Zhu, Delphine Bernhard, and Iryna Gurevych. 2010. A monolingual tree-based translation model for sentence simplification. In *Proceedings of the 23rd international conference on computational linguistics*, pages 1353–1361. Association for Computational Linguistics.