*Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020)*, pages 3484–3489
Marseille, 11–16 May 2020

# SimplifyUR: Unsupervised Lexical Text Simplification for Urdu

**Namoos Hayat Qasmi[1], Haris Bin Zia[1], Awais Athar[2], Agha Ali Raza[1]**
[1]Information Technology University, 6th Floor, Arfa Software Technology Park, Ferozepur Road, Lahore, Pakistan
[2]European Molecular Biology Laboratory, European Bioinformatics Institute (EMBL-EBI), Cambridge, CB10 1SD, UK
{namoos.qasmi, haris.zia, agha.ali.raza}@itu.edu.pk
awais@ebi.ac.uk

**Abstract**

This paper presents the first attempt at Automatic Text Simplification (ATS) for Urdu, the language of 170 million people worldwide. Being a low-resource language in terms of standard linguistic resources, recent text simplification approaches that rely on manually crafted simplified corpora or lexicons such as WordNet are not applicable to Urdu. Urdu is a morphologically rich language that requires unique considerations such as proper handling of inflectional case and honorifics. We present an unsupervised method for lexical simplification of complex Urdu text. Our method only requires plain Urdu text and makes use of word embeddings together with a set of morphological features to generate simplifications. Our system achieves a BLEU score of 80.15 and SARI score of 42.02 upon automatic evaluation on manually crafted simplified corpora. We also report results for human evaluations for correctness, grammaticality, meaning-preservation and simplicity of the output. Our code and corpus are publicly available to make our results reproducible.

**Keywords:** automatic text simplification, lexical simplification, Urdu

## 1. Introduction

Text simplification has gathered much attention recently because of its interesting applications in machine translation (Mishra et al., 2014; Štajner and Popovic, 2016) and language learning for people with learning disabilities (Carroll et al., 1998; Chen et al., 2016). Lexically complex text is also difficult to understand for children and non-native speakers (Petersen and Ostendorf, 2007; De Belder and Moens, 2010). Automatic Text Simplification (ATS) to replace lexically complex words with their simpler equivalents is thus necessary to make text understandable for a wide variety of audiences.

Urdu, the official language of Pakistan, is an Indo-Aryan language with over 170 million speakers throughout Pakistan, India and several other countries[1]. It uses Arabic script which is written from right to left. Text Simplification for Urdu exhibits a number of challenges due to its morphological richness, use of case markers and lack of linguistic resources. Urdu is a low resource language in terms of standard linguistic resources (Cieri et al., 2016). Therefore, neither data-driven approaches such as Neural Text Simplification (NTS) (Nisioi et al., 2017) that rely on complex-simplified parallel corpora, nor rule-based methods that depend on lexical resources such as WordNet[2] for generating simplifications are applicable.

Glavaš and Štajner (2015) proposed LIGHT-LS, a resource-light lexical simplification method that works well for languages with moderate morphological diversity such as English. Our approach adapts LIGHT-LS to handle challenges of Urdu text simplification by using additional features and achieves significant improvement in performance (18.4%) over the LIGHT-LS baseline. We also present a new manually constructed parallel corpus of complex-simplified Urdu sentences for benchmarking Urdu lexical simplification tasks. We have conducted both automatic and human evaluation of our proposed system and have made our code and lexical resources open-source[3] to make our results reproducible and facilitate further research on Urdu text simplification. To the best of our knowledge, this is the first ever attempt to produce the outcomes mentioned above.

## 2. Related Work

The earliest text simplification systems were rule-based and relied on lexicons like WordNet to substitute pre-defined complex words with their simpler synonyms (Bautista et al., 2009; De Belder and Moens, 2010). The main weaknesses of these systems were their low recall (De Belder and Moens, 2010) and inaccurate identification of complex words (Shardlow, 2014).

The research direction shifted from knowledge based approaches to data driven methods with the emergence of Simple Wikipedia[4]. The sentence aligned original and simple Wikipedia corpus has been used extensively in unsupervised approaches (Yatskar et al., 2010; Biran et al., 2011) and also in supervised methods (Horn et al., 2014) for text simplification.

Recent text simplification approaches treat simplification task as monolingual Machine Translation (MT) problem. Such approaches exploit statistical machine translation models, such as phrase-based machine translation (PBMT) (Štajner et al., 2015), tree-based machine translation (TBMT) (Woodsend and Lapata, 2011) or syntax-based machine translation (SBMT) (Xu et al., 2016) for text simplification. Nisioi et al. (2017) explored Neural Text Simplification (NTS) (after Neural Machine Translation (NMT)) using LSTM based encoder-decoder models and outperformed their statistical counterparts. More recently, Vu et al. (2018) used memory augmented RNNs (a.k.a Neural Semantic Encoders (NSE)) for simplification.

The applicability of above mentioned approaches is constrained to the availability of WordNets or parallel corpora, which is a barrier in low-resource settings. In

[1] https://www.ethnologue.com/language/urd

[2] Urdu WordNet does exist but contains only 6000 high frequency nouns, verb, adjectives and adverbs http://www.cle.org.pk/clestore/urduwordnet.htm

[3] https://github.com/NamoosQasmi/SimplifyUR

[4] https://simple.wikipedia.org

contrast, our approach is resource-light following Glavaš and Štajner (2015) and requires only plain text corpora.

## 3. Contribution

The specific novel contributions of this paper are:

- The first text simplification pipeline and pre-trained models (including the largest word2vec model) for Urdu.
- A novel technique to handle case-markers after simplification. Our proposed method predicts case-markers with an accuracy of 89.83% (see section 4.3).
- The first publicly available complex-simplified parallel corpus for evaluating and benchmarking Urdu text simplification tasks.

## 4. Urdu Text Simplification

Our simplification pipeline operates by identifying complex candidate words and replacing them with their simplest synonyms, selected from a pool of simplification candidates. The candidate pool is obtained from a distributional semantic model, word2vec, and candidates are ranked using several features. These steps are described in detail below.

### 4.1 Identifying Complex Candidates

In order to select the content words (words that are simplifiable), we trained a Conditional Random Field (CRF) based Parts of Speech (PoS) tagger on Urdu monolingual corpus (Jawaid et al., 2014). The corpus consists of 95.4 million tokens tagged with 41 tags and is available publicly[5]. The PoS tagger had F1 (macro) of 0.85 on independent test set. The complex sentence is first PoS tagged and nouns (NN), adjectives (ADJ), verbs (VB), adverbs (ADV) and quantifiers (Q) are selected as content words.

Next, we check content words in frequent word list of 370 Urdu words. The content words are not simplified if they occur in frequent word list.

### 4.2 Selecting Simpler Word Replacements

We employed word2vec (continuous bag-of-words) (Mikolov et al., 2013) to learn distributed representations of Urdu words. We initialized our model with the pre-trained word2vec representations (Haider, 2018) trained on three different Urdu corpora containing over 140 million tokens. To ensure diversity of genres and good representations of low frequency words, we additionally crawled a corpus of 103 million tokens from Hamariweb[6], BBC Urdu[7], Jang News[8] and Urdu Digest[9]. We trained continuous bag-of-words with a window size of 5 words to obtain 300-dimensional vector representations.

For each content word $w_i$, each semantically similar word is selected as a simplification candidate $s_i$. The similarity is computed as the cosine of angle between the vector representations of words. We select the set of top 10 most similar words as simplification candidates $(s_1,\ldots,s_{10})$ excluding the morphological derivatives of content word. Since word frequency is the most reliable predictor of word complexity and complex words tend to be rarer (Paetzold and Specia, 2016), we sort the simplification candidates in descending order by their frequency of occurrence in the Urdu corpora.

### 4.3 Verifying Grammaticality

A simplification candidate $s_i$ is only acceptable if it fits into the sequence of words preceding and following the content word $w_i$ in the complex sentence. In order to evaluate this fitness, we consider a simplification candidate only if it is predicted by the language model with the preceding and following words of the content word in the same order. In other words, if $w_{i-2}w_{i-1}w_iw_{i+1}w_{i+2}$ is the original sequence then $w_i$ is replaceable with $s_i$ only if $w_{i-2}w_{i-1}s_iw_{i+1}w_{i+2}$ is a likely sequence as per language model. We trained a trigram language model (*LM*) on 120.9 million words from above mentioned Urdu corpora and retrieve the trigrams $w_{i-2}w_{i-1}s_i$ and $s_iw_{i+1}w_{i+2}$, for each simplification candidate $s_i$.

**Handling Grammatical Case:** Case determines the grammatical function of a word in a phrase, clause or sentence. Urdu has six cases namely nominative, ergative, dative/accusative, instrumental, genitive and locative (Butt and King, 2001). In Urdu case markers can precede, follow or occur simultaneously before and after the word (see Table 1). Following Butt and King (2001), we use the case makers: کا، کے، کی، کو، کے لیے، سا، سے، سی، ساتھ، پار، پاس، پر، تک، تلک، میں، نے. One word may take multiple case markers depending on the context e.g. word کہا can be preceded by نے، سے، کو. Also, different words with similar meaning may have different case markers. This implies that case mismatch may occur while replacing words which may lead to grammatical errors. One such example is given in Table 2.

| | |
|---|---|
| ہمیں پردوں کے لیے مناسب رنگ کا انتخاب بھی کرنا ہے<br>(*We also need to choose an appropriate color for the curtains*) | a. |
| یہ رسالہ تصاویر سے مزین ہے<br>(*This magazine is adorned with pictures*) | b. |
| انہیں چلنے میں دشواری کا سامنا ہے<br>(*They face difficulty in walking*) | c. |

Table 1: Example of a case marker that (a) precedes a word (b) follows a word (c) occurs simultaneously before and after. English translations are in parenthesis and case markers are highlighted.

| | |
|---|---|
| تکلیف میں مبتلا<br>(*suffering from pain*) | a. |
| تکلیف میں شکار<br>(*victim from pain*) | b. |
| تکلیف کا شکار<br>(*victim of pain*) | c. |

Table 2: Example of case mismatch. (a) is complex and grammatically correct (b) is simplified but grammatically incorrect (c) is simplified and grammatically correct. English translations are in parenthesis.

[5] https://lindat.mff.cuni.cz/repository/xmlui/handle/11858/00-097C-0000-0023-65A9-5

[6] https://hamariweb.com/articles

[7] https://www.bbc.com/urdu

[8] https://jang.com.pk

[9] https://urdudigest.pk

There are multiple scenarios of case markers which need to be dealt during simplification. These scenarios along with examples (and corresponding English translations) are listed below.

1. Same case marker before and after replacement.

   In : دشمن اس بات سے انجان تھا

   Out : دشمن اس بات سے بے خبر تھا

   (*The enemy was unaware of this*)

2. Different case marker before and after replacement.

   In : وہ ان کا احترام کرتے ہیں

   Out : وہ ان کی عزت کرتے ہیں

   (*They respect them*)

3. No case marker before and after replacement.

   In : کشادہ ہال پورا بھر چکا تھا

   Out : وسیع ہال پورا بھر چکا تھا

   (*The spacious hall was full*)

4. No case marker before but case marker after replacement.

   In : مخاصمانہ رویہ مناسب نہیں ہے

   Out : دشمنی کا رویہ مناسب نہیں ہے

   (*Hostile behavior is not appropriate*)

5. Case marker before but no case marker after replacement.

   In : وہ یہاں کا باشندہ ہے

   Out : وہ یہاں مقیم ہے

   (*He is a resident here*)

We use the language model to predict the best possible case marker (or remove it completely) for simplification candidates. For each simplification candidate $s_i$, we retrieve the following two trigram likelihoods $w_{i\text{-}2}w_{i\text{-}1}s_i$ and $s_iw_{i+1}w_{i+2}$ from the language model. If $w_{i\text{-}2}w_{i\text{-}1}s_i$ (or $s_iw_{i+1}w_{i+2}$) is not predicted by the language model, we check if $w_{i\text{-}1}$ in $w_{i\text{-}2}w_{i\text{-}1}s_i$ (or $w_{i+1}$ in $s_iw_{i+1}w_{i+2}$) is a case marker. If not then $s_i$ is not a valid simplification, otherwise we replace $w_{i\text{-}1}$ in $w_{i\text{-}2}w_{i\text{-}1}s_i$ (or $w_{i+1}$ in $s_iw_{i+1}w_{i+2}$) with all possible case markers $c_j$ where $c_j \neq w_{i\text{-}1}$ (or $c_j \neq w_{i+1}$) and compute likelihoods for trigrams $w_{i\text{-}2}c_js_i$ (or $s_ic_jw_{i+2}$). Next, we sort the trigrams $w_{i\text{-}2}c_js_i$ (or $s_ic_jw_{i+2}$) based on their likelihood probabilities. In the rare case if no such $c_j$ exist that satisfies the trigram $w_{i\text{-}2}c_js_i$ (or $s_ic_jw_{i+2}$), we remove the case maker from the context and look for trigram $w_{i\text{-}3}w_{i\text{-}2}s_i$ (or $s_iw_{i+2}w_{i+3}$) and progress only if such trigram exists. Finally, we merge the trigrams on $s_i$ and pick the first trigram $c_js_ic_j$ (or $w_{i\text{-}2}s_ic_j$ or $c_js_iw_{i+2}$ or $w_{i\text{-}2}s_iw_{i+2}$) that exist in language model. This ensures that $s_i$ is accompanied by most probable context words (that may or may not be case markers).

The pseudo-code of our simplification algorithm is given in Algorithm 1.

## 5. Evaluation

We evaluate the performance of our lexical simplification system both automatically using standard evaluation metrics as well as manually via human judgements.

### 5.1 Automatic Evaluation

We could not find any parallel corpus of complex-simplified sentence-pairs for Urdu. Therefore, to automatically assess the output of our simplification system we hand-crafted a parallel corpus. The corpus contains 500 complex sentences and their corresponding simpler variants (at least one simple sentence for each complex sentence). The complex sentences were taken from newspapers, magazines, books and literary journals while reference simplifications were created by an expert linguist. The linguist is a native speaker and holds a doctorate degree in Urdu. A randomly selected sample comprising 10% of the corpus was manually verified by two native Urdu speakers and their inter-annotator agreement was found to be 0.90 as measured using Cohen's Kappa.

The simplification pipeline was designed in stages and after thorough analysis. A baseline system was first developed by replacing each word in the sentence with its semantically similar word (calculated using word2vec). After that we gradually added the PoS tagger, language model etc. to the pipeline. We report results at each step to show the improvement in performance after the addition of each feature (Table 3).

We evaluate our system outputs using standard evaluation metrics for text simplification (Woodsend and Lapata, 2011; Xu et al., 2016). Particularly, we use BLEU (Papineni et al., 2002) to access degree of closeness to the gold reference simplifications and SARI[10] (Xu et al., 2016) to evaluate the quality of system output. The scores are computed at corpus level. Sample complex-simplified sentence-pairs along with system outputs are given in Appendix A. A comparison of our system with state-of-the-art English ATS systems and Urdu LIGHT-LS is given in Table 4.

| **Model** | **BLEU** | **SARI** |
|---|---|---|
| word2vec | 13.06 | 16.48 |
| + pos | 50.12 | 27.73 |
| + frequent_word_list | 68.30 | 41.37 |
| + sort_by_frequency | 70.04 | 42.86 |
| + *LM* | 80.15 | 42.02 |

Table 3: Results of automatic evaluation by varying the pipeline.

[10] https://github.com/cocoxu/simplification/blob/master/SARI.py

**Algorithm 1:** SIMPLIFY($w_1w_2w_3\ldots w_n$)

```
$t_1t_2t_3 \leftarrow$ POS-Tag($w_1w_2w_3\ldots w_n$)
for $i \leftarrow 1$ to $N$ do
  if $t_i \in \{NN, ADJ, VB, ADV, Q\}$ and $w_i \notin frequent\_word\_list$ then
    $candidates \leftarrow$ most-similar($w_i$) – morphological-derivations($w_i$)
    for all $s_i \in$ sort-by-frequency($candidates$, descending=$true$) do
      if $w_{i-2}w_{i-1}s_iw_{i+1}w_{i+2} \in LM$ then
        $w_i \leftarrow s_i$
        handle-grammatical-case($w_{i-2}w_{i-1}w_iw_{i+1}w_{i+2}$)
        break
      end if
    end for
  end if
end for
return $w_1w_2w_3\ldots w_n$
```

| System | BLEU | SARI |
|---|---|---|
| Wubben et al. (2012) | 81.11 | 38.56 |
| Narayan and Gardent (2014) | 53.94 | 31.40 |
| Xu et al. (2016) | 74.44 | 41.46 |
| Nisioi et al. (2017) | 87.50 | 37.25 |
| Zhang and Lapata (2017) | 88.85 | 37.27 |
| Vu et al. (2018) | 92.02 | 36.88 |
| Glavaš and Štajner (2015) | 83.54 | 34.96 |
| Glavaš and Štajner (2015) - Urdu | 83.49 | 23.65 |
| This work | 80.15 | 42.02 |

Table 4: Comparison with state-of-the-art automatic text simplification systems.

## 5.2 Human Evaluation

Following Nisioi et al. (2017), we conducted three types of human evaluations. Three native Urdu speakers aged between 30 and 40 years with a background in Computer Science, Political Science and Fine Arts were employed to obtain human judgements. All three evaluators have studied Urdu at advanced levels in their academic career. We report inter-annotator agreements (Cohen's Kappa) between human evaluators for each assessment task.

**Correctness of Changes (C):** First, human evaluators calculate the total (*Total*) number of changes made by the system. Changes are marked correct (*Correct*) if they preserve meaning and grammaticality and make sentence simpler at the same time. *Total* changes were 525 out of which 247 (average of three evaluators) were marked as *Correct*. The average inter-annotator agreement turned out to be 0.89.

**Grammaticality (G) and Meaning Preservation (M):** Second, human evaluators rate each sentence with at least one change for grammaticality and meaning preservation on a Likert scale of 1-5 with (1) being very bad and (5) being very good. The average scores were 4.75 for G and 4.30 for M and inter-annotator agreements were 0.98 for G and 0.95 for M.

**Simplicity of Sentences (S):** Third, human evaluators were shown a pair of source (complex) and target (system output) sentence and asked whether target sentence is: +2: much more simple; +1: simple; 0: equally difficult; -1: difficult; -2: much more difficult, than the source sentence. The average rank was +0.96 and inter-annotator agreement was 0.90.

## 6. Conclusion & Future Work

This paper presents the first lexical text simplification system for Urdu that efficiently handles case markers and requires only a regular corpus. We also present a parallel corpus of complex-simplified sentence-pairs for evaluating and benchmarking text simplification tasks for Urdu. Our best model achieves a BLEU score of 80.15 and a SARI score of 42.02 on automatic evaluation. We have made our code and lexical resources publicly available to facilitate further research.

Our proposed system simplifies only single words, whereas, multi-word expressions are common in Urdu. This is due to the absence of clear word boundary marker (Zia et al., 2018). For example, the word خاطرخواہ is mostly written as multi-word expression خاطر خواہ and therefore, word2vec fail to learn its representation. In the future, we plan to overcome segmentation challenges to support simplification of multi-word expressions. Furthermore, we also plan to replace trigram language model with neural language model to efficiently handle long distance dependencies.

## 7. Bibliographical References

Bautista, S., Gervás, P., & Madrid, R. I. (2009, May). Feasibility Analysis for SemiAutomatic Conversion of Text to Improve Readability. In ICTA (pp. 33-40).

Biran, O., Brody, S., & Elhadad, N. (2011, June). Putting it simply: a context-aware approach to lexical simplification. In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers-Volume 2 (pp. 496-501). Association for Computational Linguistics.

Butt, M., & King, T. H. (2001, December). Non-nominative subjects in Urdu: A computational analysis. In Proceedings of the International Symposium on Non-nominative Subjects (pp. 525-548).

Carroll, J., Minnen, G., Canning, Y., Devlin, S., & Tait, J. (1998, July). Practical simplification of English newspaper text to assist aphasic readers. In Proceedings of the AAAI-98 Workshop on Integrating Artificial Intelligence and Assistive Technology (pp. 7-10).

Chen, P., Rochford, J., Kennedy, D. N., Djamasbi, S., Fay, P., & Scott, W. (2016). Automatic text simplification for people with intellectual disabilities. Artificial Intelligence Science and Technology.

Cieri, C., Maxwell, M., Strassel, S., & Tracey, J. (2016, May). Selection criteria for low resource language programs. In Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16) (pp. 4543-4549).

De Belder, J., & Moens, M. F. (2010). Text simplification for children. In Prroceedings of the SIGIR workshop on accessible search systems (pp. 19-26). ACM; New York.

Glavaš, G., & Štajner, S. (2015, July). Simplifying lexical simplification: Do we need simplified corpora?. In Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers) (pp. 63-68).

Haider, S. (2018, May). Urdu Word Embeddings. In Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018).

Horn, C., Manduca, C., & Kauchak, D. (2014, June). Learning a lexical simplifier using wikipedia. In Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers) (pp. 458-463).

Jawaid, B., Kamran, A., & Bojar, O. (2014, May). A Tagged Corpus and a Tagger for Urdu. In LREC (pp. 2938-2943).

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781.

Mishra, K., Soni, A., Sharma, R., & Sharma, D. (2014, August). Exploring the effects of sentence simplification on Hindi to English machine translation system. In Proceedings of the Workshop on Automatic Text Simplification-Methods and Applications in the Multilingual Society (ATS-MA 2014) (pp. 21-29).

Narayan, S., & Gardent, C. (2014, June). Hybrid simplification using deep semantics and machine translation.

Nisioi, S., Štajner, S., Ponzetto, S. P., & Dinu, L. P. (2017, July). Exploring neural text simplification models. In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers) (pp. 85-91).

Paetzold, G., & Specia, L. (2016, June). Semeval 2016 task 11: Complex word identification. In Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016) (pp. 560-569).

Papineni, K., Roukos, S., Ward, T., & Zhu, W. J. (2002, July). BLEU: a method for automatic evaluation of machine translation. In Proceedings of the 40th annual meeting on association for computational linguistics (pp. 311-318). Association for Computational Linguistics.

Petersen, S. E., & Ostendorf, M. (2007). Text simplification for language learners: a corpus analysis. In Workshop on Speech and Language Technology in Education.

Shardlow, M. (2014, May). Out in the Open: Finding and Categorising Errors in the Lexical Simplification Pipeline. In LREC (pp. 1583-1590).

Štajner, S., Béchara, H., & Saggion, H. (2015, July). A deeper exploration of the standard PB-SMT approach to text simplification and its evaluation. In Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers) (pp. 823-828).

Štajner, S., & Popovic, M. (2016). Can text simplification help machine translation?. In Proceedings of the 19th Annual Conference of the European Association for Machine Translation (pp. 230-242).

Vu, T., Hu, B., Munkhdalai, T., & Yu, H. (2018). Sentence simplification with memory-augmented neural networks. arXiv preprint arXiv:1804.07445.

Woodsend, K., & Lapata, M. (2011, July). Learning to simplify sentences with quasi-synchronous grammar and integer programming. In Proceedings of the conference on empirical methods in natural language processing (pp. 409-420). Association for Computational Linguistics.

Wubben, S., Van Den Bosch, A., & Krahmer, E. (2012, July). Sentence simplification by monolingual machine translation. In Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1 (pp. 1015-1024). Association for Computational Linguistics.

Xu, W., Napoles, C., Pavlick, E., Chen, Q., & Callison-Burch, C. (2016). Optimizing statistical machine translation for text simplification. Transactions of the Association for Computational Linguistics, 4, 401-415.

Yatskar, M., Pang, B., Danescu-Niculescu-Mizil, C., & Lee, L. (2010, June). 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 (pp. 365-368). Association for Computational Linguistics.

Zhang, X., & Lapata, M. (2017). Sentence simplification with deep reinforcement learning. arXiv preprint arXiv:1703.10931.

Zia, H. B., Raza, A. A., & Athar, A. (2018). Urdu Word Segmentation using Conditional Random Fields (CRFs). arXiv preprint arXiv:1806.05432.

## Appendix A: Example System Outputs

| | | |
|---|---|---|
| 1 | a | موٹرکار نظروں سے اوجھل ہو گئی |
| | b | موٹرکار نظروں سے چھپ گئی<br>موٹرکار نظروں سے غائب ہو گئی |
| | c | موٹرکار نظروں سے غائب ہو گئی |
| | d | The motorcar disappeared from sight |
| 2 | a | وہ سب کی توجہ کا مہور ہے |
| | b | وہ سب کی دلچسپی کا مرکز ہے |
| | c | وہ سب کی دلچسپی کا مرکز ہے |
| | d | He is everyone's center of attention |
| 3 | a | انسان جسمانی اور دماغی تھکاوٹ میں مبتلا ہے |
| | b | انسان جسمانی اور دماغی تھکن کا شکار ہے |
| | c | انسان جسمانی اور دماغی تھکن کا شکار ہے |
| | d | Man is suffering from physical and mental fatigue |
| 4 | a | ان میں سے کتنے لوگ ہیں جو وقت آنے پر بہت ہمت اور شجاعت کا مظاہرہ کرتے ہیں |
| | b | ان میں سے کتنے لوگ ہیں جو وقت آنے پر بہت حوصلے اور بہادری کا مظاہرہ کرتے ہیں |
| | c | ان میں سے ہزاروں لوگ ہیں جو وقت آنے پر بہت ہمت اور بہادری کا اظہار کرتے ہیں |
| | d | How many of them are ones who show great courage and bravery when the time comes |
| 5 | a | قوس افراد کے لیے سیر محض دلچسپی نہیں ضرورت ہے |
| | b | قوس افراد کے لیے سیر صرف دلچسپی نہیں ضرورت ہے |
| | c | سنبلہ افراد کے لیے سیر محض دلچسپی نہیں ضرورت ہے |
| | d | Outing is not merely an interest but a need for Sagittarius individuals |

Table A: Examples of (a) complex sentence (b) gold reference(s) and (c) system output (d) English translation. Complex words are highlighted orange and simplifications are highlighted green (if correct) and red (if incorrect).