Creating a Corpus Resource for Text Simplification R & D

Poster · November 2017

DOI: 10.13140/RG.2.2.22257.38241

5 authors, including:

Debra Revere
University of Washington Seattle
94 PUBLICATIONS **1,265** CITATIONS
SEE PROFILE

Partha Mukherjee
The Pennsylvania State University, Great Valley, Malvern, PA, USA
48 PUBLICATIONS **327** CITATIONS
SEE PROFILE

David Kauchak
Middlebury College
48 PUBLICATIONS **1,338** CITATIONS
SEE PROFILE

Gondy Leroy
The University of Arizona
165 PUBLICATIONS **2,535** CITATIONS
SEE PROFILE

Creating a Corpus Resource for Text Simplification R & D

Debra Revere, MLIS, MA[1], Partha Mukherjee, PhD[2], David Kauchak, PhD[3], Gondy Leroy, PhD[2]

[1]Univ of WA, Seattle, WA; [2]Univ of AZ, Tucson, AZ; [3]Pomona College, Claremont, CA

Introduction

Text simplification (TS) involves rewriting complex text into simpler language that is easier to understand while retaining meaning. An important TS resource is text in which the difficulty of the text is known. Previous corpora have either relied on the text source (e.g., blogs vs. medical articles) or human perception to judge difficulty. These assessments have some value, but are subjective. To better quantify text difficulty, we have created a collection of texts and questions (Qs) which test knowledge (As) within those texts in order to systematically quantify text difficulty. Our Q/A process combines auto-generated and human-annotated methods to create a training corpus from texts of varying difficulty within the medical domain. To minimize factors that might influence performance other than text difficulty, we standardized the process using Heilman's (2011) framework which parses each sentence and then uses a rule-based approach to generate Qs based on syntactic structure[1].

Text Simplification Corpus Creation

Following auto Q/A generation, annotators received one Q/A file for 60 medical conditions that contained: 1) auto-generated Qs in Wh- (i.e., What, When, etc.) and True/False formats; 2) the A embedded in its complete original text; 3) the A; and 4) a score which indicated viability of the text to answer to the Q. Files with scores lower than 2.5; illegible characters (e.g., HajduGÇôCheney syndrome); or non-medical topics were eliminated—leaving a total of 8,637 Q/A item sets across 56 articles. Human annotators randomly selected 36 articles on common (e.g., dyslexia) and uncommon (e.g., Spondylocostal dysostosis) topics to ensure a wide variety of language and complexity. Annotators corrected grammatical errors and conducted up to 3 manual simplifications per Q/A item set to create a Q/A set that was simpler but retained lexical structure—syntactically reducing complexity by splitting sentences, eliminating non-standard abbreviations, or rephrasing strings containing double negatives. Table 1 provides annotation examples.

Table 1. Auto-generated Q/A item set input and human annotated Q/A item set output examples.

Auto Q/A	Is the disease related to the SRRT gene?	The disease is related to the SRRT
Annotated Q/A	What disease is related to the SRRT gene?	Spondylocostal dysplasia
Auto Q/A	What is acute rheumatic fever rare in in children?	In children, acute rheumatic fever is rare in most of the developed world.
Annotated Q/A	In most of the developed world is prevalence of acute rheumatic fever rare in children?	YES

Results

The final training corpus contains 1264 questions: 670 Wh-Qs with 4 annotated multiple-choice As (1 correct, 3 incorrect) and 594 True/False-Qs with YES, NO, Sometimes and Not enough information As.

Conclusion

This corpus is the foundation for future TS work in which we will create smaller text fragments with matching Qs at different difficulty levels and generate comprehension data for readers with varying levels of literacy.

Acknowledgments: Research reported in this poster was supported by the National Library of Medicine of the National Institutes of Health under award number R01LM011975. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

References

1. Heilman M. Automatic factual question generation from text. (PhD Dissertation: CMU-LTI-11-004). Pittsburgh PA: Carnegie Mellon Univ; 2011.