Reproducing a Manual Evaluation of the Simplicity of Text Simplification System Outputs

Maja Popović, Sheila Castilho, Rudali Huidrom, Anya Belz
ADAPT Centre @ Dublin City University
{name.surname}@adaptcentre.ie

Engaging Content
Engaging People

Text simplification

- transform uncommon words or long and complicated sentences
  - retain the meaning
  - make it easier to read by people/process by computers
- (human) evaluation of ATS systems:
  - Meaning Preservation (how much of the original meaning is retained in the output)
  - Grammaticality (whether the grammar of the generated output is good)
  - Simplicity (how difficult/simple the generated output is)

this paper focuses on evaluation of simplicity

Experiments

Original experiment

- the first attempt of using neural networks for ATS
- two basic neural text simplification (NTS) variants:
  - NTS: relying only on internal word representations
  - NTS-W2V: additionally ising external word2vec representations
- each system variant generated outputs in three ways:
  - NTS-DEFAULT, NTS-W2V-DEFAULT, : beam search with size 5
  - NTS-BLEU, NTS-W2V-BLEU: re-ranking an n-best list using the automatic metric BLEU
  - NTS-SARI, NTS-W2V-SARI: re-ranking using the automatic metric SARI
- compared with three publicly available ATS outputs generated in previous work
  - PBSMT: phrase-based SMT system with re-ranking
  - textscSARI+PPDB: system based on paraphrasing and SARI score
  - LIGHTLS: lexical simplification based on word representations

- all nine systems evaluated
  - **manually** for meaning preservation, grammaticality and **simplicity**
  - automatically by the BLEU metric and the SARI metric

Data

- 359 publicly available sentences originating from English Wikipedia
- all simplified by the nine ATS systems
- all evaluated automatically
- first 70 sentences evaluated manually
  (9 systems → 630 sentences in total)

Evaluation of simplicity

- three non-native English speakers
- presented with the original sentence and an automatically generated simplification of it
- asked to assign a score to each pair according to the following guidelines:
  - +2 if the simplified version is much simpler than the original,
  - +1 if the simplified version is somewhat simpler than the original,
  - 0 if they are equally simple/difficult,
  - -1 if the simplified version is somewhat more difficult than the original, and
  - -2 if the simplified version is much more difficult than the original.

- aggregated system-level scores are reported in the paper (mean sentence-level scores)
- indicate that NTS model substantially outperform all of the previous systems in terms of simplicity

Reproduction experiment

- same data as the original one
- three different non-native speakers
- the same instructions

! further details were not available

Details known only for the reproduction study:

- ***Native languages of evaluators***
  each evaluator had a different native language (Serbian, Brazilian Portuguese and Manipuri)
- ***Evaluators' background***
  all the evaluators were computational linguistics researchers
- ***Evaluators' experience with TS and its evaluation***
  one evaluator had experience with TS evaluation, the other two did not;
  they needed a few additional instructions and examples to fully understand
  the concept of simplicity
  and to be able to separate it from meaning and grammar
- ***Number of sentences assessed by each evaluator***
  the experienced evaluator annotated all sentences
  the other two evaluators annotated half of the sentences each
- ***Number of multiply annotated sentences used for IAA***
  each sentence was annotated by two evaluators
  IAA is computed on the whole set

Results

Comparing different ATS systems

automatic text simplification system	Simplicity original rank	Simplicity original score	Simplicity reproduction rank	Simplicity reproduction score	small-sample coefficient of variation (CV*) ↓
NTS DEFAULT	(3)	0.46	(5)	0.33	5.41
NTS-SARI	(5)	0.38	(3/4)	0.34	1.69
NTS-BLEU	(1)	0.92	(3/4)	0.34	22.0
NTS-W2V-DEFAULT	(6)	0.21	(6)	0.32	4.84
NTS-W2V-SARI	(2)	0.63	(1)	0.46	6.66
NTS-W2V-BLEU	(4)	0.40	(2)	0.36	1.68
PBSMT	(9)	-0.55	(7)	0.08	35.6
SARI+PPDB	(7)	0.03	(9)	0.01	0.99
LIGHTLS	(8)	-0.01	(8)	0.03	1.98

- coefficient of variation CV* as measure of reproducibility
  - indicate that for some systems human scores are more reproducible than for others
  - however, not obvious why this is the case
- the main claim from the original paper is confirmed
  (NTS generates simpler outputs than previous systems)
  ! different tendencies regarding comparison of individual NTS systems
- correlation between the original and the reproduced results:
  - Pearson's r: 0.766 (moderate to high)
  - Spearman's ρ: 0.787

Inter-annotator agreement

quadratic Cohen's Kappa:

- original: 0.66
- reproduced: 0.40

- difference is hard to interpret due to missing information about the original experiment:
  - what sub-set of sentences was used for IAA
  - how many annotators per sentence
  - evaluators' experience with TS and the notion of simplicity

- possible factors for lower IAA in the reproduction study:
  - only one evaluator had experience with TS
  - this evaluator annotated the entire test set
    → IAA only between experienced and inexperienced annotators
- however, only a speculation

⇒ availability of the sentence-level scores from the original study would have helped

Comparison with reproducing automatic scores

metric	output	evaluation round original	repr1	repr2	repr3	repr4	CV* ↓
BLEU ↑	NTS default	84.51	84.50	85.60	84.20	–	0.838
(automatic)	NTS-w2v default	87.50	–	89.36	88.80	–	1.314
SARI ↑	NTS default	30.65	30.65	30.65	–	–	0
(automatic)	NTS-v2w default	31.11	–	31.11	–	–	0
Simplicity ↑	NTS default	0.46	–	–	–	0.33	5.41
(human)	NTS-v2w default	0.21	–	–	–	0.32	4.84

- 'original' = results reported in the original paper
- 'repr1' = results reported in an earlier reproduction paper (Cooper and Shardlow, 2020)
- 'repr2' and 'repr3" = results reported in recent reproduction paper (Belz et al. 2022)
- 'repr4' = results from this work

  - CV values higher for human evaluation than for automatic scores

⇒ human evaluation was more difficult to reproduce

Conclusions

A general tendency regarding human evaluations:

- details about human evaluation process available only in papers dealing with human evaluation itself
- no such details in papers where human evaluation is only a method to assess systems/models
- even if the models and/or outputs are made publicly available, human evaluations are not

- probable reasons
  - these details are not considered important
  - conditions were not optimal ⇒ fear of negative reviews
    - small portion of text evaluated
    - small number of evaluators participated
    - very small (or none) portion of text evaluated by more than one evaluator for IAA

Our recommendations:

- for authors: always report all the details;
  providing them is more scientifically useful than no information for fear of negative reviews
- for reviewers: do not penalise human evaluations carried out in sub-optimal conditions

Acknowledgements

The ADAPT Centre is funded by Science Foundation Ireland through the SFI Research Centres Programme and is co-funded under the European Regional Development Fund (ERDF) through Grant 13/RC/2106.