Linguistic Capabilities for a Checklist-based evaluation in Automatic Text Simplification

Oscar M. Cumbicus-Pineda[1,3], Itziar Gonzalez-Dios[2] and Aitor Soroa[2]

[1]*Ixa group, University of the Basque Country (UPV/EHU)*

[2]*Ixa group, HiTZ center, University of the Basque Country (UPV/EHU), Informatika Fakultatea, Manuel Lardizabal 1, 20018 Donostia*

[3]*Carrera de Computación, Facultad de la Energía las Industrias y los Recursos Naturales No Renovables, Universidad Nacional de Loja, Loja, Ecuador*

Abstract

Evaluation in Automatic Text Simplification (ATS) has been carried out by means of automatic metrics such as SARI, BLEU, by manual analysis that takes into account the grammar/fluency, meaning preservation and simplicity of the outputs, readability metrics or by extrinsic evaluation via NLP tasks. These metrics and dimensions give an overview of what the systems are doing, but we do not exactly which are the strong and weak points. Inspired by recent literature of Natural Language Processing tasks for classification, in this paper we explore the checklist-based evaluation of the linguistic capabilities ATS systems need to meet. We apply this evaluation to a syntax aware edit-based ATS system and we point out which are the weakness and the strength of the system, which can also lead to improvements of the system.

Keywords

Automatic Text Simplification, Manual evaluation, Checklists, Capabilities

1. Introduction

Automatic Text Simplification is a Natural Language Processing (NLP) research line which aims to reduce the complexity of a text at both lexical and syntactic levels for a certain target audience. The interested reader is referred to the following works for detailed information about ATS [1, 2, 3, 4, 5, 6].

As with many other Natural Language Generation (NLG) tasks [7], the evaluation of ATS is still an open question that arises big concerns in the community. Automatic metrics do not capture all the nuances of text simplification, and human evaluation is costly and difficult to reproduce. Still, the research community is putting a lot of effort on systematising ATS evaluation and making manual evaluation as reliable as possible [8, 9]. Moreover, it is worth to notice that most of the works on ATS simplification focus exclusively on English.

The most successful methods in ATS today are based on deep learning techniques, and are often cast as a machine translation task where the system learns to "translate" from complex sentences

Proceedings of the First Workshop on Current Trends in Text Simplification (CTTS 2021), co-located with SEPLN 2021. September 21st, 2021 (Online). Saggion, H., Štajner, S. and Ferrés, D. (Eds).

ocumbicus001@ikasle.ehu.es and oscar.cumbicus@unl.edu.ec (O. M. Cumbicus-Pineda); itziar.gonzalezd@ehu.eus (I. Gonzalez-Dios); a.soroa@ehu.eus (A. Soroa)

0000-0001-5483-0913 (O. M. Cumbicus-Pineda); 0000-0003-1048-5403 (I. Gonzalez-Dios); 0000-0001-8573-2654 (A. Soroa)

to simpler counterparts. Evaluation of neural models is, however, a difficult task, as they are black boxes that are trained in an end-to-end fashion. Current trends in evaluating and debugging neural models are focusing on methods and metrics that go beyond the traditional metrics (e.g. accuracy) such as adversarial rules (perturbations of the input by preserving its semantics e.g changing 'what' to 'which', 'movie' to 'film' or introducing a typo, but inducing changes in a black box model's predictions) [10] or checklists (a matrix of general linguistic capabilities and tests) [11]. For example, in the case of sentiment analysis tasks, the identification of words that carry positive, negative, or neutral sentiment, comparatives and superlatives, negation or named entities are the linguistic characteristics considered. Although checklists are primarily intended for classification tasks, in our opinion, they can also be applied to generation tasks such as ATS.

In this paper, we open a path towards the study of the linguistic capabilities required for ATS with the aim of better understanding how ATS systems work. This way, we know which are the weak and strong spots of the systems. To that end, we analyse the outputs of three different neural ATS systems trained in English, Italian and Spanish and we present the evaluation of a system. The contributions of this paper are: i) analysis of the outputs of three different systems for three languages, ii) a list of linguistic capabilities for ATS, iii) a checklist evaluation of a system and future directions to improve it.

This paper is structured as follows: in Section 2 we present the approaches to evaluate ATS systems, in Section 3 we detail our approach and describe the linguistic capabilities, in Section 4 we present the checklist based evaluation of a system and we conclude and outline the future work in Section 5.

2. Evaluation in ATS

Evaluation in ATS is a research concern for the community. At the moment, systems are usually evaluated automatically or evaluated via human ratings.

Regarding the automatic evaluation, the most used automatic metrics are BLEU [12], and SARI [13]. These metrics are language independent since they mainly rely on n-gram overlap (BLEU) or measuring the words that are added, delete or kept (SARI). Other metrics, however, need language dependent tools such as a parser in the case of SAMSA [14] or a question generation and answering systems as in QUESTEVAL for Sentence Simplification [15]. However, some of these tools are not available for many languages and cannot be applied. Readability assessment metrics such as Flesch–Kincaid [16] are also language-dependent, in this case, for English. Although some readability metrics have been adapted to some languages e.g. Fernandez-Huerta index for Spanish [17], not all the languages have their own formulae. Other metrics that have been used and proposed to evaluate ATS systems are TER [18], ROUGE [19], C-Score [20] or the E-Score [21]. To ease the process of calculating automatic metrics and facilitating comparison, the package EASSE was created [22], which includes BLEU, SARI and Flesch-Kincaid Grade Level.

Concerning human evaluation, three criteria are mainly used: grammar/fluency, meaning preservation, and simplicity [5]. Commonly, a Likert scale from 1-5 is used to give the ratings and in the Quality Assessment for Text Simplification shared-task, a tree level scale was used with the bad/ok/good levels for each of the criteria [23]. They also created a combination of the three scores called overall, which rewarded more meaning preservation and simplicity than grammaticality.

Human judgments, however, can vary across the target audience of the simplification and the evaluators. Moreover, the evaluators can be linguists, simplification experts, members of the target audience or crowdsourcing workers. They all can be paid or not. In order to assist evaluators, a reading comprehension test was done in [24] and specific questions for the task were posed in [25]. In this study, the authors also asked evaluators about the original sentences, since in the dataset they cured, depending on the language, more than 15 % of the sentences were not correct.

Other techniques that have been used to evaluate ATS systems are information measures against a specially curated reference corpus to evaluate different linguistic phenomena [26], eye-tracking [27] or extrinsic evaluation via information extraction [28, 29], a chunk-based question generation system [25], machine translation [30, 31], or semantic role labelling [29].

As Alva-Manchego et al. [5] point out, metrics such as BLEU and SARI are flawed, and it is necessary to keep all their limitations in mind. Regarding the human evaluation criteria (leaving apart the costs and possible bias) they wonder if grammar/fluency, meaning preservation, and simplicity are enough. Moreover, we do not know what is happening and what neural systems *understand*, unless an error analysis is made. In this line, Shardlow and Nawaz [32] present a framework of six types of error found in clinical neural ATS. These error types can be summarised as changes with or without loss or alteration of the original meaning, reduction of the information leading the miss of critical information, word repetitions and no changes.

In order to better understand what systems (not restricted to neural) are simplifying, in this paper we propose a checklist evaluation for ATS by focusing on linguistic and simplicity phenomena or capabilities. Checklist and similar techniques have been successfully used in other NLP tasks such as sentiment analysis, duplicate question detection, machine comprehension [11], contradiction detection in dialogue [33], hate speech detection [34], offensive content detection [35], or bias analysis [36]. Some capabilities such as the negation have also been studied [37] across different natural language inference tasks. By using general linguistic capabilities, our aim is also to be as language independent as possible. We are aware that this evaluation is also expensive, but it is necessary to understand and find the weak spots of the systems, which can lead to the development of methods to improve them. To our knowledge, this is also the first time checklists are proposed for generation tasks.

3. Checklist-based Evaluation: Linguistic Capabilities for ATS

To create the list of the capabilities, we have analysed the outputs of three ATS neural systems: a re-implementation of the edit-based system EditNTS [38] (EditNTS), a syntax aware edit-based system [39] (Edit+Synt), and transformer built by us. We have trained and tested these systems in the following corpora: for English, in Wikilarge/TurkCorpus [40, 41, 42], for Spanish in Simplext [43] and for Italian in the combination of a subset of the PaCCSS-it corpus [44], the SIMPITIKI corpus [45], the Terence-Teacher corpus [46]. In Table 1 we show the results of the systems for each dataset.

We have randomly selected a sample of 15 original sentence pairs for each dataset, together with the outputs of each system. In Table 2 we show an example of a sentence from Wikilarge. Based on this sample, we have analysed the features related to grammar, meaning preservation and simplicity included in the sentences and we have created a list of them. We have also identified

	EditNTS		Edit+Synt		Transformer	
	SARI	**BLEU**	**SARI**	**BLEU**	**SARI**	**BLEU**
Wikilarge	36.75	72.99	**36.97**	**75.35**	33.27	41.66
Simplext	36.52	7.31	**39.48**	**7.51**	35.30	0.30
PaCCSS-it-SIMPITIKI-TerenceTeacher	51.95	53.43	**52.25**	**53.54**	33.56	18.68

Table 1
Results of the analysed systems

Type	Output
Complex	It is situated at the coast of the Baltic Sea, where it encloses the city of Stralsund.
Simple (manual)	It is situated at the coast of the Baltic Sea. It encloses the city of Stralsund.
EditNTS	it is situated at the coast of the baltic sea, where it is the city of stralsund.
Edit+Synt	it is situated at the coast of the baltic sea. it is it also encloses the city of stralsund.
Transformer	it of old old old old old old germanic people.

Table 2
Example of the outputs

important features not related to these dimensions. This analysis has been carried out by a linguist expert on text simplification, native in Spanish and with C1 proficiency in English and Italian. We have decided to analyse only 15 sentences for each language because we realised that the most important errors were repeating.

3.1. Capabilities for ATS

In the following sections, we present the linguistic capabilities we propose to reveal the strong and weak points of ATS systems. These capabilities are meant to be useful for general simplification, but they can be adapted depending on the target audience and the purpose of simplification. As mentioned before, ATS is manually evaluated on the basis of three dimensions: grammar/fluency, meaning preservation and simplicity. We add two new dimensions to this list: the *prerequisites*, which lists a set of basic checklists any simplification should comply with, and *ethical aspects*, which measure any ethical issue that may arise because of the produced simplifications. We define these capabilities in terms of general linguistic phenomena so that they can be applied to all the languages.

3.1.1. Prerequisites

Two types of prerequisites are needed to check before the manual evaluation starts. The first one is the **no simplification** (**P0**), which means that the original sentence does not need to be simplified and, therefore, the output of the simplified sentences should be a copy of the input e.g. *Take the square root of the variance.*. In this case, the evaluation does not continue. If the system, however, does not simplify a sentence that needs to be simplified, the capability is not satisfied and the evaluation is stopped, because the system has simply copied the original one.

The second one is related to **system errors** and these sentences should be discarded since they are not fluent enough to be evaluated. Based on the most common system errors, the capabilities we propose are:

- **(P1)** No UNK tokens: *el de UNK y los UNK.*
- **(P2)** No non required quotation marks or strange characters: *the ” ” ” ” ” ” ” ” ” ” ” ” ” ” ” ”*
- **(P3)** No relation or alignments problems: *While at Kahn he was chief architect for the Fisher Building in 1928 . -> he was the current conductor of the boston symphony orchestra.*

3.1.2. Grammar/Fluency

These capabilities take into account the structure of the language. Although it has been argued that the fluency of neural systems is close to human fluency, this has been only validated for English [47]. That is why fluency and grammar need to be taken into account when evaluating ATS outputs.

- Word level:
  - **(G1)** No repeated words e.g. repeated determinants (*las las casas*), adjectives (*the same whole whole was*), nouns, prepositions, or conjugated verbs (*estos fueron fue*)...
  - **(G2)** No tense change, unless necessary for a certain target audience (*sarebbe -> è*) or modality (*è -> puó essere*) in the verbs
- Morpho-syntactic level:
  - **(G3)** Correct and finished phrases or sentences (∗*against their own,* ∗*defensores de los derechos humanos en,* ∗*en ahora,* ∗*. is the national parks' biggest* or ∗*promete aumentar que las misiones en 2010*)
  - **(G4)** Correct agreement of subject and verb (*he volunteers (...) and* ∗*search*), grammatical gender of determinants, nouns and adjectives (∗*dentro de las apoyo* or ∗*a organization*), contractions (∗*de el* or ∗*can not*)...
  - **(G5)** Correct phraseology: verb + preposition (∗*enviar en los habitantes*), collocations...
  - **(G6)** No repeated arguments: double subject or verb (∗*it is it also encloses*)...
- Cohesion level (both at inter- and intra-sentence levels):
  - **(G7)** Correct punctuation (*as , , he*)
  - **(G8)** Correct grammatical order: phrase order, sentence order
  - **(G9)** Correct coreference
  - **(G10)** No Definiteness change (*the organization -> an organization*)

3.1.3. Meaning preservation

Meaning preservation is related to the fidelity of the simplified sentence in relation with the original sentence in terms of the meaning. As the capabilities may differ from sentence to sentence e.g. presence of negation or adverbial information, we propose two levels in this dimension: the mandatory capabilities, which should be always evaluated, and the optional capabilities, which will be evaluated only in the cases where the phenomena is present e.g. negation should be evaluated if there is a negation in the original sentence.

- Mandatory
  - **(M1)** Important information kept (all the arguments and adjuncts that are necessary to understand the whole meaning of the sentence).

  - **(M2)** Register (formal, informal, literary, technical...) kept, unless required by the target audience
  - **(M3)** No meaning change or only subtle nuances changes e.g. deleting or adding emphasizers (*la risposta è tecnicamente no.*)
- Optional (depending on the sentence)
  - **(M4)** Named entities unaltered (*La ministra de Defensa -> la ministra de asuntos sociales*)
  - **(M5)** Negation kept
  - **(M6)** Temporal adverbs and relations kept
  - **(M7)** Numerical expressions kept or/and not altered except for rounding (check simplicity capabilities [48])
  - **(M8)** Correct lexical simplifications (*project focuses on the laws of motion-> project focuses on health care*)
  - **(M9)** No too general lexical simplification (*educated workers -> people*)
  - **(M10)** No unnecessary cliches, idioms that affect the meaning (*Ma non è tutto ! -> ma non è tutto oro quel che luccica.*)

3.1.4. Simplicity

These capabilities are related to simplification studies and guidelines [49], and to summaries of easy-to-read guidelines [50]. As in the meaning dimension we also define here mandatory and optional capabilities.

Mandatory
- **(S1)** Shorter sentences (explanations should be added in our opinion in another sentence)
- **(S2)** Same term for same concept
- **(S3)** Logical or temporal ordering of relations
- **(S4)** Active voice (instead of passive)
- **(S5)** Simple, frequent words
- **(S6)** Same term consistently used
- **(S7)** Only one main idea per sentence covered
- **(S8)** Only one finite verb for sentence
- **(S9)** Simple punctuation

Optional
- **(S10)** No legal, foreign and technical jargon
- **(S11)** 'you' used to speak directly to readers
- **(S11)** Use of the number and not the word
- **(S13)** Rounded numerical expressions
- **(S14)** More known names for named entities
- **(S15)** Necessary and correct elaborations, explanations
- **(S16)** Elided arguments or verbs recovered
- **(S17)** No exceptions to exception

3.1.5. Ethical aspects

The research and analysis of ethical aspects has gained a lot of importance in the last years in NLP. Given that one of ATS' goals is to adapt texts to people with difficulties, special care should be taken and the maxims *Primum non nocere* or *do no harm* should be of a great importance. That is why we think that these dimensions should also be taken into account. Following we present two ethical *violations* we have found in our analysis.
- **(E1)** No wrong information or misinformation (*Disney received a full-size Oscar statuette and seven miniature ones, presented to him by 10-year-old child actress Shirley Temple. -> Disney sold*

to him by in old child shirley temple.), unnecessary/wrong elaborations (*in the area Provence-Alpes-Côte Azur in the Nord-Pas-de-Calais region.*), hallucinations or explanations/ information which we do not know if they are true or not (*Military career Donaldson enlisted in the Australian Army on 18 June 2002 . -> War war II military career Donaldson left the united states on 18 june 2002.*)

- **(E2)** No non-present stereotypes or unnecessary mentions to discriminate/minoritary groups: *Detenidos tres menores por amenazar e injuriar a otra menor a través de una red social -> la guardia civil detiene a a los red emigrantes.*

To our knowledge, ethical aspects have not be taken into account in the evaluation of ATS and we are open to discuss them with the community, as well as with other capabilities.

3.2. Capability score

In order to quantitatively evaluate the aforementioned capabilities, we propose to score each capability separately for each sentence. So, for each sentence, the evaluator indicates whether a capability has been fulfilled with a binary score (1:yes, 0:no). For example, the sentence simplified by Edit+Synt presented in Table 2 misses the capability G6 at the grammatical dimension and S5 in the simplicity dimension.

To score a sample/corpus, we calculate the percentage of the positive scores for each capability. That is, if we are evaluating the capability G1 in a sample of 50 sentences, and it is fulfilled in 48 of them, the score of the capability G1 will be 96 %. In the case of the optional capabilities, only the sentences that have that feature should be taken into account.

To interpret the scores we propose a scale (Table 3) with the following values: 96-100 % perfect, 81-95 % substantial, 61-80 % moderate and $<$ 60 % low. This scale is inspired by the interpretation of Cohen's kappa, but, being the one of the main aims of ATS help people to understand texts, we think that we need to be hard with the rating and that is why all the capabilities below 60 % are considered low. The capabilities that also score less than 80 % should be addressed by system developers.

Score	Interpretation
$<$ 60 %	Low
61-80 %	Moderate
81-95 %	Substantial
96-100 %	Perfect

Table 3
Interpretation of the capability scale values

4. Case study: Checklist evaluation for Edit+Synt at Wikilarge

As case study, in this section we evaluate the capabilities of the sentences simplified by Edit+Synt (system with best quantitative performance) in the Wikilarge corpus. We have randomly chosen 10 % of the test set (36 sentences) to carry out this analysis (the ones used to create the list of capabilities were discarded). 5 of the sentences were sentences where no simplification should

be carried out and no simplification was performed. So, we have annotated in total 31 sentences according to the the capabilities. The annotator is an expert on text simplification and, once trained in the task, she spent an average of 90 seconds per sentence pair. The annotation was done in a spreadsheet. In Table 4 we group the capabilities by their score. We only show the optional capabilities if there are 5 more sentence to evaluate.

Score	Capabilities
Low (< 60 %)	P0, G3, M1, M3, M8, M9, S1, S7, S8, S10
Moderate (61-80 %)	G4, S4, S5
Substantial (81-95 %)	G1, G2, G5, G6, G7, G8, M2, M4, S3, S9, E1, E2
Perfect (96-100 %)	P1, P2, P3, G9, G10, M7, S2, S6

Table 4
Results of the mandatory capabilities of Edit+Synt at Wikilarge

Let us explain the results by grouped dimension (In table 5 we present the examples of the violated capabilities to illustrate the errors.). In the case of the prerequisites, the ones related to the systems errors (P1, P2, and P3) are successfully fulfilled. However, in the case of no simplification (P0), the system has not simplified 9 sentences (which, as mentioned before, were discarded from evaluation), but only two of them were correctly unaltered. This result suggests that preprocessing should be applied before performing any simplification step, as proposed by Scarton et al. [51].

Regarding the grammar, the weakest points of the system are related to the correction of the phrases (G3) and the agreements (G4). This indicates that the system fails to properly exploit phrase level information. Strong points of the system are, however, word (G1 and G2) and cohesion (G7, G8, G9 and G10) level capabilities.

With respect to the meaning preservation, the system struggles to keep important information (M1), meaning changes (M3), and lexical simplifications (M8 and M9). This indicates that the system tends to perform too many delete operations that affect the original message of the sentence. Lexical simplification is also a challenge for this system, which can be addressed by performing it on its own as many other systems do. The strong points are related to the register (M2), and the preservation of named entities (M4) and numerical expressions (M7).

Concerning the simplicity, the systems really struggles to create shorter sentences (S1), and simple sentences (S7 and S8). This is linked to the splitting operation, which is a challenge for deep learning based systems. In order to overcome this problem, performing rule based simplification as a previous step has been proposed [52]. Moreover, the system does not handle technical words (S10) and does not use simple frequent words (S4), most likely due to the fact that theses words are missing in the training data. This could be addressed together with the lexical simplification. Some effort should also put to avoid passive voice (S5). The strong capabilities are, however, the use of the same term for the same concept (S2) and the perseverance of logical and temporal ordering (S3). This may be due to the fact that the system does not perform any reordering operation and follows the order of the original sentence.

Finally, looking at the ethical aspects, we see that the system has not added stereotypes (E2) (% 93,5), but it is close to the limit regarding the addition of wrong information (% 83,9) (E1). The sentences that contain added wrong information or hallucinations tend to be longer than the

Violated capability	Complex	Output
P0	He was appointed Companion of Honour (CH) in 1988.	He was appointed Companion of Honour (CH) in 1988. **the name was given to the nobel prize in physics.**
G3	There he had one daughter, later baptized as Mary Ann Fisher Power, to Ann (e) Power.	He had one daughter, **later ∅ as Mary Ann Fisher Power.**
G4	The wounds inflicted by a club are generally known as bludgeoning or blunt-force trauma injuries.	**The wounds was** by a club are generally known as bludgeoning or blunt-force .
M1	They are rivaled as biological materials in toughness only by chitin.	they are only ∅ by chitin.
M3	The vagina is remarkably elastic and stretches to many times its normal diameter during vaginal birth.	The vagina is very thin elastic and stretches to many times its normal diameter during ∅ birth.
M8	The tongue is sticky because of the presence of glycoprotein-rich mucous, which both lubricates movement in and out of the snout and helps to catch ants and termites, which adhere to it.	The tongue is sticky because of the presence of glycoprotein-rich mucous, which both lubricates movement. Out of the snout and helps to catch ants and termites, which **brings** to it.
M9	Stands were eventually added behind each set of goals during the 1980s and 1990s as the ground began to be modernised.	Stands were eventually added behind each set of goals during the 1980s and 1990s as the ground began to be **finished.**
S1/S7/S8	Civil unrest in northern Italy spawns the medieval musical form of Geisslerlieder, penitential songs sung by wandering bands of Flagellants.	civil unrest in northern Italy made the medieval musical form of Geisslerlieder, **penitential songs sung by wandering bands of Flagellants.**
E1	He left a detachment of 11,000 troops to garrison the newly conquered region.	He left a detachment of 11,000 troops to garrison the newly conquered region **around the world, and was used in the area.**

Table 5
Examples of violated capabilities

original ones, so this can be a hint to detect this kind of errors.

In general, we can say that the system passes the *checklist exam*, since many capabilities are in the ranges of perfect and substantial. However, there are weak points that should be addressed and treated, which we know thanks to this methodology.

5. Conclusion and Future Work

In this paper we have proposed to evaluate the ATS systems based on their linguistic capabilities. To that end, inspired by the checklist method, we have defined a first set of linguistic capabilities

required so that a sentence/text be a correct simplification. These capabilities are grouped in the three dimensions manual ATS is usually evaluated but we have also added the prerequisites and the ethical aspects. We think that adding the ethical aspects is important since one of the main aims of ATS is help people to understand texts and no wrong information or biases should be included or amplified. Moreover, based on these capabilities, we can understand what systems are doing and which are their weak and strong points systematically. This can lead to open ways to improve the systems and future research.

We also have proven the validity of the proposal to evaluate ATS systems by analysing the outputs of the Edit+Synt system. Based on this evaluation, we have seen that the system performs quite well but needs improvements in the correction of phrases and agreement, keeping the important information, creating shorter sentences.

We are open to discuss more capabilities with the community, adapt them or specify them. We would like to test other languages, other systems and even to automatise the analysis of some of the features to facilitate the manual evaluation. As suggested by the reviewers, it will also be interesting to i) stratify the analysis sample of the datasets to e.g based on sentence length and depth, readability measures to analyse other kind of errors and create more capabilities; ii) carry out the analysis in other dataset with other domains that can include abstract language, figurative language, and sarcasm; iii) perform pilot studies to determine a better threshold for the interpretation of the capability score, carry out analysis in other dataset with other domains that can include abstract language, figurative language, and sarcasm and; iv) explore how to visualise the evaluation; and, finally, v) compare our results to the ones obtained with a traditional (human and automatic) evaluation method and try to find correlations.

There is a lot of work to do until we get outputs that can be used by people that adapt and/or simplify texts or directly by people who need the simplified/adapted texts. In this sense, checklist evaluation of linguistic capabilities can open a way towards a better quality of ATS.

Acknowledgments

We really thank the anonymous reviewers for their comments and suggestions. We acknowledge the following projects: DeepText (KK-2020/00088), DeepReading RTI2018-096846-B-C21 (MCIU/AEI/FEDER, UE), BigKnowledge for Text Mining, BBVA and IXA group (Basque Government (excellence research group IT1343-19).

References

[1] I. Gonzalez-Dios, M. J. Aranzabe, A. Díaz de Ilarraza, Testuen sinplifikazio automatikoa: arloaren egungo egoera, Linguamática 5 (2013) 43–63.

[2] M. Shardlow, A Survey of Automated Text Simplification, International Journal of Advanced Computer Science and Applications 4 (2014) 58–70.

[3] A. Siddharthan, A Survey of Research on Text Simplification, ITL-International Journal of Applied Linguistics 165 (2014) 259–298.

[4] H. Saggion, Automatic Text Simplification, Synthesis Lectures on Human Language Technologies 10 (2017) 1–137.

[5] F. Alva-Manchego, C. Scarton, L. Specia, Data-driven sentence simplification: Survey and benchmark, Computational Linguistics 46 (2020) 135–187.
[6] P. Sikka, M. Singh, A. Pink, V. Mago, A Survey on Text Simplification, arXiv preprint arXiv:2008.08612 (2020).
[7] A. Celikyilmaz, E. Clark, J. Gao, Evaluation of text generation: A survey, arXiv preprint arXiv:2006.14799 (2020).
[8] A. Belz, S. Agarwal, Y. Graham, E. Reiter, A. Shimorina (Eds.), Proceedings of the Workshop on Human Evaluation of NLP Systems (HumEval), Association for Computational Linguistics, Online, 2021. URL: https://www.aclweb.org/anthology/2021.humeval-1.0.
[9] A. Shimorina, A. Belz, The human evaluation datasheet 1.0: A template for recording details of human evaluation experiments in nlp, 2021. arXiv:2103.09710.
[10] M. T. Ribeiro, S. Singh, C. Guestrin, Semantically equivalent adversarial rules for debugging NLP models, in: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2018, pp. 856–865.
[11] M. T. Ribeiro, T. Wu, C. Guestrin, S. Singh, Beyond accuracy: Behavioral testing of NLP models with CheckList, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Online, 2020, pp. 4902–4912. URL: https://www.aclweb.org/anthology/2020.acl-main.442. doi:10.18653/v1/2020.acl-main.442.
[12] K. Papineni, S. Roukos, T. Ward, W.-J. Zhu, Bleu: a method for automatic evaluation of machine translation, in: Proceedings of the 40th annual meeting of the Association for Computational Linguistics, 2002, pp. 311–318.
[13] W. Xu, C. Napoles, E. Pavlick, Q. Chen, C. Callison-Burch, Optimizing statistical machine translation for text simplification, Transactions of the Association for Computational Linguistics 4 (2016) 401–415. URL: https://www.aclweb.org/anthology/Q16-1029. doi:10.1162/tacl_a_00107.
[14] E. Sulem, O. Abend, A. Rappoport, Semantic Structural Evaluation for Text Simplification, in: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), 2018, pp. 685–696.
[15] T. Scialom, L. Martin, J. Staiano, E. Villemonte de la Clergerie, B. Sagot, Rethinking Automatic Evaluation in Sentence Simplification, 2021. arXiv:2104.07560.
[16] J. P. Kincaid, R. P. Fishburne Jr, R. L. Rogers, B. S. Chissom, Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel, Technical Report, Naval Technical Training Command Millington TN Research Branch, 1975.
[17] J. Fernández Huerta, Medidas sencillas de lecturabilidad, Consigna 214 (1959) 29–32.
[18] M. Snover, B. Dorr, R. Schwartz, L. Micciulla, J. Makhoul, A study of translation edit rate with targeted human annotation, in: Proceedings of association for machine translation in the Americas, volume 200, Citeseer, 2006.
[19] C.-Y. Lin, Rouge: A package for automatic evaluation of summaries, in: Text summarization branches out, 2004, pp. 74–81.
[20] I. Temnikova, G. Maneva, The C-Score–Proposing a Reading Comprehension Metrics as a Common Evaluation Measure for Text Simplification, in: Proceedings of the Second

Workshop on Predicting and Improving Text Readability for Target Reader Populations, 2013, pp. 20–29.

[21] S. Mathias, P. Bhattacharyya, How Hard Can it Be? The E-Score-A Scoring Metric to Assess the Complexity of Text, in: Proceedings of Quality Assessment for Text Simplification (QATS) Workshop, 2016, pp. 10–14.

[22] F. Alva-Manchego, L. Martin, C. Scarton, L. Specia, Easse: Easier automatic sentence simplification evaluation, in: EMNLP-IJCNLP 2019-Conference on Empirical Methods in Natural Language Processing and 9th International Joint Conference on Natural Language Processing (demo session), 2019, pp. 49–54.

[23] S. Štajner, M. Popovic, H. Saggion, L. Specia, M. Fishel, Shared Task on Quality Assessment for Text Simplification, in: Proceedings of the Quality Assessment for Text Simplification (QATS), 2016, pp. 22–31.

[24] A. A. Mandya, T. Nomoto, A. Siddharthan, Lexico-syntactic text simplification and compression with typed dependencies, in: 25th International Conference on Computational Linguistics, 2014.

[25] I. Gonzalez-Dios, M. J. Aranzabe, A. D. de Ilarraza, Making Biographical Data in Wikipedia Readable: A Pattern-based Multilingual Approach, in: Proceedings of the Workshop on Automatic Text Simplification-Methods and Applications in the Multilingual Society (ATS-MA 2014), 2014, pp. 11–20.

[26] C. Gasperin, E. Maziero, S. M. Aluisio, Challenging choices for text simplification, in: International Conference on Computational Processing of the Portuguese Language, Springer, 2010, pp. 40–50.

[27] L. Rello, S. Bautista, R. Baeza-Yates, P. Gervás, R. Hervás, H. Saggion, One half or 50%? An eye-tracking study of number representation readability, in: IFIP Conference on Human-Computer Interaction, Springer, 2013, pp. 229–245.

[28] R. J. Evans, Comparing Methods for the Syntactic Simplification of Sentences in Information Extraction, Literary and Linguistic Computing 26 (2011).

[29] R. Evans, C. Orasan, Sentence Simplification for Semantic Role Labelling and Information Extraction, in: Proceedings of Recent Advances in Natural Language Processing, 2019, p. 285–294.

[30] K. Mishra, A. Soni, R. Sharma, D. M. Sharma, 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), 2014, pp. 21–29.

[31] S. Štajner, M. Popović, Can text simplification help machine translation?, in: Proceedings of the 19th Annual Conference of the European Association for Machine Translation, 2016, pp. 230–242.

[32] M. Shardlow, R. Nawaz, Neural text simplification of clinical letters with a domain specific phrase table, in: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, 2019, pp. 380–389.

[33] Y. Nie, M. Williamson, M. Bansal, D. Kiela, J. Weston, I like fish, especially dolphins: Addressing Contradictions in Dialogue Modelling, arXiv preprint arXiv:2012.13391 (2020).

[34] P. Röttger, B. Vidgen, D. Nguyen, Z. Waseem, H. Margetts, J. Pierrehumbert, Hatecheck: Functional tests for hate speech detection models, arXiv preprint arXiv:2012.15606 (2020).

[35] S. Bhatt, R. Jain, S. Dandapat, S. Sitaram, A case study of efficacy and challenges in practical human-in-loop evaluation of nlp systems using checklist, in: Proceedings of the Workshop on Human Evaluation of NLP Systems (HumEval), 2021, pp. 120–130.

[36] L. Li, X. Chen, H. Ye, Z. Bi, S. Deng, N. Zhang, H. Chen, On robustness and bias analysis of bert-based relation extraction, arXiv e-prints (2020) arXiv–2009.

[37] M. M. Hossain, V. Kovatchev, P. Dutta, T. Kao, E. Wei, E. Blanco, An analysis of natural language inference benchmarks through the lens of negation, in: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2020, pp. 9106–9118.

[38] Y. Dong, Z. Li, M. Rezagholizadeh, J. C. K. Cheung, EditNTS: An neural programmer-interpreter model for sentence simplification through explicit editing, in: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Florence, Italy, 2019, pp. 3393–3402. URL: https://www.aclweb.org/anthology/P19-1331. doi:10.18653/v1/P19-1331.

[39] O. M. Cumbicus-Pineda, I. Gonzalez-Dios, A. Soroa, A Syntax-Aware Edit-based System for Text Simplification, in: Proceedings of RANLP 2021, 2021.

[40] Z. Zhu, D. Bernhard, I. Gurevych, A monolingual tree-based translation model for sentence simplification, in: Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010), 2010, pp. 1353–1361.

[41] X. Zhang, M. Lapata, Sentence simplification with deep reinforcement learning, in: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, 2017, pp. 584–594.

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

[43] H. Saggion, S. Štajner, S. Bott, S. Mille, L. Rello, B. Drndarevic, Making it simplext: Implementation and evaluation of a text simplification system for spanish, ACM Transactions on Accessible Computing (TACCESS) 6 (2015) 1–36.

[44] D. Brunato, A. Cimino, F. Dell'Orletta, G. Venturi, 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, 2016, pp. 351–361.

[45] S. Tonelli, A. P. Aprosio, F. Saltori, Simpitiki: a simplification corpus for italian, Proc. of CLiC-it (2016).

[46] D. Brunato, F. Dell'Orletta, G. Venturi, S. Montemagni, Design and annotation of the first Italian corpus for text simplification, in: Proceedings of The 9th Linguistic Annotation Workshop, Association for Computational Linguistics, Denver, Colorado, USA, 2015, pp. 31–41. URL: https://www.aclweb.org/anthology/W15-1604. doi:10.3115/v1/W15-1604.

[47] L. Martin, A. Fan, É. de la Clergerie, A. Bordes, B. Sagot, Multilingual unsupervised sentence simplification, arXiv preprint arXiv:2005.00352 (version 16 Apr 2021) (2020).

[48] S. Bautista, R. Hervás, P. Gervás, R. Power, S. Williams, A system for the simplification of numerical expressions at different levels of understandability, in: Natural Language Processing for Improving Textual Accessibility (NLP4ITA 2013), 2013, pp. 10–19.

[49] I. Gonzalez-Dios, M. J. Aranzabe, A. D. de Ilarraza, The corpus of Basque simplified texts (CBST), Language Resources and Evaluation 52 (2018) 217–247.

[50] R. Mitkov, S. Štajner, The fewer, the better? a contrastive study about ways to simplify,

in: Proceedings of the Workshop on Automatic Text Simplification - Methods and Applications in the Multilingual Society (ATS-MA 2014), Association for Computational Linguistics and Dublin City University, Dublin, Ireland, 2014, pp. 30–40. URL: https://www.aclweb.org/anthology/W14-5604. doi:10.3115/v1/W14-5604.

[51] C. Scarton, P. Madhyastha, L. Specia, Deciding when, how and for whom to simplify, in: ECAI 2020, volume 325, IOS Press, 2020, pp. 2172–2179.

[52] M. Maddela, F. Alva-Manchego, W. Xu, Controllable Text Simplification with Explicit Paraphrasing, arXiv preprint arXiv:2010.11004 (2020).