Sentence Simplification Capabilities of Transfer-Based Models

Sanja Štajner[1], Kim Cheng Sheang[2], Horacio Saggion[2]
[1] Symanto Research, Nuremberg, Germany
[2] LaSTUS Lab / TALN Group, Universitat Pompeu Fabra, Spain
stajner.sanja@gmail.com, kimcheng.sheang@upf.edu, horacio.saggion@upf.edu

Abstract

According to the official adult literacy report conducted in 24 highly-developed countries, more than 50% adults, on average, can only understand basic vocabulary, short sentences, and basic syntactic constructions. Everyday information found in news articles is thus inaccessible to many people, impeding their social inclusion and informed decision-making. Systems for automatic sentence simplification aim to provide scalable solution to this problem. In this paper, we propose new state-of-the-art sentence simplification systems for English and Spanish, and specifications for expert evaluation that are in accordance with well-established easy-to-read guidelines. We conduct expert evaluation of our new systems and the previous state-of-the-art systems for English and Spanish, and discuss strengths and weaknesses of each of them. Finally, we draw conclusions about the capabilities of the state-of-the-art sentence simplification systems and give some directions for future research.

Introduction

To be able to make informed decisions and actively participate in society, people need to understand written information, especially up-to-date information such as news. Yet, the results of adult (age 16–65) literacy report (OECD 2013) which involved 24 countries[1] revealed that as much as 16.7% of the population from those countries, on average, only understands basic vocabulary, and approximately 50% understands only basic syntactic constructions. Those numbers are even higher for some countries, e.g. in the US, 21.7% of the population only understands basic vocabulary, while in Spain, 28.3% of the population understands only basic vocabulary and 67.4% of the population understands only basic syntactic constructions (OECD 2013). Non-native speakers, and people with various reading or intellectual impairments, also have problems understanding lexically and syntactically complex sentences (Carroll et al. 1998; Aluísio et al. 2008; Saggion et al. 2015; Orăsan, Evans, and Mitkov 2018).

Since the late nineties, many initiatives raised awareness about fundamental information being written in a way that is too difficult to understand for many people. They proposed guidelines for how to write more accessible texts (Nomura, Nielsen, and Tronbacke 1997; Freyhoff et al. 1998; Mencap 2002; Karreman, van der Geest, and Buursink 2007; W3C 2008; Cooper et al. 2010; PlainLanguage 2011). Websites offering accessible information exist in many countries (Štajner and Saggion 2018), but they depend on well-trained human editors, and can thus offer only a handful of articles at the time. This problem attracted the attention of natural language processing (NLP) community, and created the task of Automatic Text Simplification (ATS) (Carroll et al. 1998; Saggion 2017).

Due to the high social impact the field can make, the emergence of parallel (original-simple) text simplification corpora, neural architectures, and large pretrained language models, the field started attracting significantly more attention in the last several years (Alva-Manchego, Scarton, and Specia 2020; Štajner 2021). As opposed to the earlier rule-based sentence simplification models, e.g. (Siddharthan 2006; Saggion et al. 2015; Ferrés et al. 2016), which require considerable amount of handcrafted rules by linguistic experts, the neural architectures are more straightforward to train if large amounts of training data and sufficient computational power are available. However, as they do not directly target any particular simplification operations (unlike the rule-based systems), it is much harder to know whether the simplifications they perform create easy-to-read texts or not. The output of those systems is commonly evaluated by several automatic measures that calculate its similarity to the original sentence and the 'gold standard' manually simplified ones, and by crowdsourced evaluations of its grammaticality, simplicity, and meaning preservation (Alva-Manchego, Scarton, and Specia 2020). Those types of evaluation have several shortcomings, and do not assess whether or not the output follows easy-to-read guidelines (Štajner 2021).

To fill those gaps in this important research area, we make the following contributions:

- We propose new transformer-based sentence simplifica-

[1] Participating countries: Australia, Austria, Belgium (Flanders), Canada, the Czech Republic, Denmark, Estonia, Finland, France, Germany, Ireland, Italy, Japan, Korea, the Netherlands, Norway, Poland, the Slovak Republic, Spain, Sweden, the United Kingdom (England and Northern Ireland), the United States, Cyprus, and the Russian Federation.

Type	Problem	Solution
Lexical	low frequency	replacement by more frequent words and phrases
	lengthy words	replacement by shorter words
Syntactic	long sentence	sentence splitting or removal of non-essential information
	apposition	sentence splitting or removal of apposition (if non-essential)
	relative clause	sentence splitting or removal of subordinate clause (if non-essential)

Table 1: Linguistic obstacles and simplification operations to remove them.

tion systems for English and Spanish that, according to an extensive multi-facet evaluation, show state-of-the-art performances for both languages.[2]

- We propose guidelines for expert human evaluation that rely on guidelines for producing easy-to-read texts. They thus better reflect potential usability for real target readers than the traditionally performed crowdsourced human evaluation that relies on subjective feelings of simplicity by non-expert evaluators.
- We conduct extensive expert evaluations of several state-of-the-art sentence simplification systems for English and Spanish. This allows us to better assess usefulness of those systems in real-world scenarios.
- We pinpoint the strengths and weaknesses of each of the state-of-the-art systems for sentence simplification in English and Spanish, and suggest ways to improve them for achieving better social impact.

Related Work

Linguistic Obstacles to Text Comprehension

Some of the guidelines for how to write more accessible texts are more detailed than others, e.g. the Plain Language guidelines (PlainLanguage 2011) are more detailed than “Make it simple” guidelines (Freyhoff et al. 1998) and “Am I making myself clear?” (Mencap 2002). Nevertheless, they all share the same basic concepts: write short sentences, use the simplest form of a verb (present tense and not conditional or future), use short and simple words, avoid unnecessary words, information and cross references, etc. In automatic sentence simplification, those guidelines motivate simplification operations that need to be performed, e.g. sentence splitting, deletion of non-essential sentence parts, lexical simplification by using shorter and more frequent words. The most frequently addressed problems and solutions are given in Table 1. Although complex discourse structures and presence of non-essential information at text level are known to play a major role in making texts difficult to understand (Kintsch and van Dijk 1978; Štajner and Hulpuş 2018), most ATS systems operate only at sentence level, and rarely perform any removal of non-essential information (Alva-Manchego, Scarton, and Specia 2020). Rule-based sentence simplification systems (Siddharthan and Mandya 2014; Saggion et al. 2015; Ferrés et al. 2016) perform specific transformations, e.g. removing appositions, creating separate sentences from relative clauses, or removing non-essential information. Data-driven approaches, in contrast, learn from transformations found in training data, and do not target any particular linguistic obstacles. They are often very conservative, either leaving sentences unchanged or performing only one or two isolated lexical simplifications and occasional sentence splitting (Štajner and Nisioi 2018).

Evaluation of Sentence Simplification Systems

Automatic sentence simplification systems are usually evaluated in two ways: (1) automatically, for the similarity of their output to the gold standard manual simplifications; and (2) manually, for grammaticality, simplicity, and meaning preservation of their output sentences.

For automatic evaluation, studies commonly use ‘gold standard’ manually simplified test sentences, and calculate the BLEU (Papineni et al. 2002) and SARI (Xu et al. 2016) scores. Automatic evaluation is useful for quickly getting rough estimates of performances of different system configurations. Nevertheless, although both scores show some correlations with human assessments (Štajner, Mitkov, and Saggion 2014; Xu, Callison-Burch, and Napoles 2015), they are not reliable enough for comparing performances of different simplification systems (Sulem, Abend, and Rappoport 2018; Vásquez-Rodríguez et al. 2021). Some studies also use Flesch-Kincaid Readability Index (Flesch 1949) for automatic evaluation. Although well-known in readability research, this metric is considered inadequate for sentence simplification (Saggion 2017; Štajner 2021; Tanprasert and Kauchak 2021).

In the ideal scenario, grammaticality and meaning preservation should be evaluated by native speakers with high literacy levels, as the original sentences can be too complex to understand for an average reader (see Introduction). Simplicity, in contrast, should be evaluated by non-native speakers, experts in text simplification or production of easy-to-read texts, or carers of the intended target population (Štajner 2021). All three evaluations are usually performed using a five point Likert scale (Alva-Manchego, Scarton, and Specia 2020). This type of evaluation often has the following shortcomings: (1) in most of the studies, all three evaluations are performed by the same people, usually Amazon Mechanical Turk workers, whose literacy levels are unknown, and who thus might not be the optimal evaluators of grammaticality and meaning preservation; (2) if the pool of evaluators is comprised of mixture of native and non-native speakers, or people with different literacy levels, the notion of simplicity and grammaticality might differ among them.

[2] The code and data is available at https://github.com/KimChengSHEANG/TS-AAAI_2022

State-of-the-Art Sentence Simplification Systems

Specia (2010) was the first to address data-driven sentence simplification as a monolinguial machine translation (MT) problem, translating from original to simple Brazilian Portuguese. This approach was adopted by many subsequent studies which attempted at English sentence simplification by using phrase-based MT (Coster and Kauchak 2011; Wubben, van den Bosch, and Krahmer 2012), syntax-based MT (Zhu, Bernhard, and Gurevych 2010; Xu et al. 2016), or neural MT (Nisioi et al. 2017; Zhang and Lapata 2017). A detailed manual evaluation showed that neural MT model (Nisioi et al. 2017) outperforms the phrase-based MT model (Wubben, van den Bosch, and Krahmer 2012) and the syntax-based MT model (Xu et al. 2016), by producing more grammatical and simpler outputs, while better preserving the original meaning (Nisioi et al. 2017; Štajner and Nisioi 2018). Apart from performing lexical simplifications, the system proposed by Nisioi et al. (2017) made several correct sentence shortenings and sentence splittings (Štajner and Nisioi 2018). It was found (Štajner and Nisioi 2018) that the system performs more sentence shortenings when trained on English Wikipedia (Hwang et al. 2015) than when trained on English Newsela dataset (Xu, Callison-Burch, and Napoles 2015), and that it performs sentence splitting only if trained on Newsela dataset. Those findings show that the system's simplification capabilities heavily depend on the transformations present in the training data.

Zhao et al. (2018) introduced transformer-based sentence simplification by using the original Transformer model (Vaswani et al. 2017), and integrating the simple paraphrase database (Pavlick and Callison-Burch 2016) into it. The performances of several versions of their system were compared with the earlier systems only by using automatic evaluation metrics (Zhao et al. 2018), thus not offering a clear picture of system's sentence simplification capabilities.

Scarton and Specia (2018) demonstrated that adding an artificial control token at the beginning of the original English sentences in encoder-decoder architectures with attention leads to better simplification output according to automatic evaluation measures. They experimented with control tokens that encode the desired grade level of the output, desired simplification operation, or both. Several subsequent studies used various control tokens in either unsupervised (Kariuk and Karamshuk 2020; Martin et al. 2021), or supervised (Martin et al. 2020, 2021; Sheang and Saggion 2021) neural sentence simplification. All sentence simplification systems that use control tokens were evaluated using automatic evaluation metrics (BLUE and SARI). The three supervised systems: ACCESS (Martin et al. 2020), MUSS (Martin et al. 2021), and the system proposed by Sheang and Saggion (2021), were additionally evaluated via crowdsourced annotations of grammaticality, simplicity, and meaning preservation (Martin et al. 2021; Sheang and Saggion 2021). However, none of the systems was analysed for the type and correctness of the transformations, nor named entity hallucinations and disappearances.

In this study, we fill those gaps by proposing three new transfer-based sentence simplification systems with control tokens and performing extensive expert human evaluations to compare the new systems to the previous state of the art.

Experiments

Models

In this work, we use three transformer-based models:

- **mBart** (Liu et al. 2020): a multilingual sequence-to-sequence model based on BART (Lewis et al. 2020), trained as a denoising auto-encoder, using random span masking and sentence shuffling on a subset of 25 languages from XLM-R dataset (Conneau et al. 2020).
- **T5** (Raffel et al. 2020): an encoder-decoder model pre-trained on multiple tasks: unsupervised tasks such as BERT-style span masking (Devlin et al. 2019), and supervised tasks such as machine translation, document summarization, question answering, classification tasks, and reading comprehension. T5 is trained on Colossal Clean Crawled Corpus (C4), a dataset created by applying a set of filters to English texts sourced from the public Common Crawl web scrape.
- **mT5** (Xue et al. 2021): a multilingual model based on T5 (Raffel et al. 2020) trained on the multilingual colossal dataset (mC4), a dataset with over 100 languages also extracted from public Common Crawl web scrape.

The T5 and mT5 models are available in different sizes, depending on the number of attention modules and the number of parameters. Due to the memory limitations and time constraints, we are able only to use T5-base for English. For Spanish, we use mT5-base and mT5-large. We implement the models using Huggingface Transformers library[3] (Wolf et al. 2020) with PyTorch[4] and Pytorch lightning.[5] Motivated by the recent studies on controllable sentence simplification outlined in the previous section, we add four control tokens to our models, previously proposed by Martin et al. (2020):

- **C**: the ratio between the length of the source and target sentences, where the length is expressed in characters.
- **L**: normalized character-level Levenshtein similarity (Levenshtein 1965) between the source and target sentences.
- **WR**: the ratio between the word rank of the target and source sentences, where the word rank stands for inverse frequency order of all words in a sentence.
- **DTD**: the ratio between the maximum depth of the dependency tree of the source and the target.

The first control token (C) controls the compression level during simplification, the second token (L) controls the level of modifications performed, the third token (WR) controls the lexical complexity at word level, and the fourth token (DTD) controls the syntactic complexity of the sentence (Martin et al. 2020). We set as the optimal values of the control tokens those values that lead to the highest SARI score

[3] https://huggingface.co/transformers/model_doc/t5.html

[4] https://pytorch.org

[5] https://pytorchlightning.ai

on the validation datasets. The search for the optimal values of control tokens is done using Optuna (Akiba et al. 2019), for each language and dataset separately.

Training Details

For T5, we perform hyperparameters search using Optuna (Akiba et al. 2019) on T5-small and reduced size of the dataset to speed up the process. We train all models with the same hyperparameters: a batch size of 6 for T5-base, 256 for maximal length in tokens, learning rate of 3e-4, weight decay of 0.1, Adam's epsilon of 1e-8, 5 warm up steps, 5 epochs. We use the rest of the parameters with their default values from Transformers library, and set the seed to 12 for reproducibility. For generation, we use beam size of 8. We train and evaluate all models using Google Colab Pro, which has a random GPU T4 or P100. Both have 16GB of memory, up to 25GB of RAM, and a time limit of 24h. Training the T5-base model for 5 epochs takes around 20 hours. For mBART and mT5, we follow the same process, and use the same hyperparameters as for T5. We train mBART and mT5 on our own computer due to the limited resources of Google Colab. We only change the batch size to adapt with the GPU memory. Our computer has Intel core i9 8950HK, 32GB of memory, and NVidia RTX 3090 GPU (24GB of memory).

Datasets

For English, we use Wiki-Large (Zhang and Lapata 2017) dataset for training. Wiki-Large is the largest and most commonly used dataset for English sentence simplification. It contains 296,402 sentence pairs from automatically-aligned complex-simple sentence pairs from document-aligned English Wikipedia and Simple English Wikipedia articles. For validation and testing in English, we use two datasets: MTurk (Horn, Manduca, and Kauchak 2014), and ASSET (Alva-Manchego et al. 2020). Both datasets contain 2,000 samples for validation and 359 samples for testing. In MTurk dataset, each sample contains an original sentence from English Wikipedia and eight simplifications of that sentence by eight Amazon Mechanical Turk workers. In ASSET, each sample contains an original sentence from English Wikipedia and ten manual simplifications. The original sentences are the same in both datasets.

For Spanish, we use automatically-aligned sentence pairs (Štajner et al. 2018) from the original and manually simplified Newsela corpus which comprises original news articles, manually simplified to several simpler levels by professional editors. The complex-simple sentence pairs were aligned using the CATS tool (Štajner et al. 2017, 2018) build especially for that purpose.[6] As the alignments of sentences between further-apart complexity levels are less reliable (Štajner et al. 2018), we only use the alignments between the original articles and the first level of simplification. The correctness of these alignments is estimated to 96.1%, for the recommended C3G sentence-level alignment (Štajner et al. 2018). From all aligned sentence pairs, we randomly select 700 sentence pairs for validation, and 350 for testing. The rest (7,414 sentence pairs), we use for training.

Rule	Simpler form...
1	Uses active tense instead of passive
2	Uses the simplest form of the verb (simple present or past tense instead of conditionals or future)
3	Avoids hidden verbs (i.e. verbs converted into a noun)
4	Avoids abbreviations
5	Uses shorter and/or more commonly used words
6	Omits unnecessary words
7	Uses the same term consistently
8	Avoids legal, technical, or foreign jargon
9	Simplifies punctuation
10	Makes the sentence(s) shorter
11	Keeps subject, verb, and object close together
12	Avoids double negatives and exceptions to exceptions
13	Places the main idea before the exceptions and conditions
14	Covers only one main idea per sentence
15	Avoids figures of speech and metaphors
16	Uses number instead of word

Table 2: Guidelines for expert annotation, based on the Plain Language guidelines (PlainLanguage 2011), "Make it simple" guidelines (Freyhoff et al. 1998), and "Am I making myself clear?" guidelines (Mencap 2002).

Standard Evaluation

To compare our systems with a larger number of previously proposed systems, we use the SARI metric (Xu et al. 2016) implemented in EASSE (Alva-Manchego et al. 2019), a simplification evaluation library. SARI compares system outputs to the references and the source sentence by counting words that are added, deleted and kept.

To compare our best systems with best previous systems (according to SARI) with different architectures, we perform crowdsourced human evaluation of grammaticality (G), meaning preservation (M), and simplicity (S) on a 1–5 Likert scale, by five Amazon Mechanical Turk workers who are native speakers of the respective language (English or Spanish). We follow the same procedure as in other studies that perform this type of evaluation, e.g. (Martin et al. 2021). The annotators are first provided with the consent form, and then the instructions and instances for evaluation. For each instance, they are provided with the original sentence and the three simplified versions. For each simplified version, they are asked to judge how much they agree (1–strongly disagree, 5–strongly agree) with the following statements (used to assess G, M, and S, respectively):

- The sentence is grammatically correct and well-formed.
- The sentence has the same meaning as the original one.
- The sentence is simpler than the original one.

In the cases where simplified version consists of several sentences, the annotators are instructed to take into account all sentences that comprise the simplified version.

[6] https://github.com/neosyon/SimpTextAlign

Transformations	Original	Automatically Simplified	System
lexical, **splitting**, **addition** ('In 1943,')	Graham **attended** Wheaton College from 1939 to 1943**, when** he **graduated with a BA** in anthropology.	Graham **went** to Wheaton College from 1939 to 1943**. In 1943,** he **got a degree** in anthropology.	T5-base
lexical, **splitting** (*missing information*)	Graham **attended** Wheaton College from 1939 to 1943**, when** he graduated *with a BA in anthropology*.	Graham **went** to Wheaton College from 1939 to 1943**. H**e graduated from Wheaton College in 1943.	MUSS-sup
addition (','), **lexical-phrase**	In **1987** Wexler was **inducted into** the Rock and Roll Hall of Fame.	In **1987,** Wexler was **added to** the Rock and Roll Hall of Fame.	T5-base
reordering, *lexical* (*missing information*)	**In 1987** *Wexler* was inducted into the Rock and Roll Hall of Fame.	*He* was inducted into the Rock and Roll Hall of Fame **in 1987**.	MUSS-sup

Table 3: Automatic English sentence simplification performed by our system (T5-base) versus the state of the art (MUSS-sup). Correct transformations are marked in bold, whereas incorrect transformations (lost or changed meaning) are marked in italics.

Score	Definition
1	Simplified sentence is meaningless.
2	Simplified sentence has completely different meaning from the original.
3	Meaning has not been changed but some essential information is missing.
4	Meaning is almost the same; there are some minor differences that are not essential.
5	Meaning is fully kept (some nuances might have been lost due to deletion of non-essential information).

Table 4: Definition of meaning preservation scores in expert evaluation.

Score	According to the rules in Table 2...
1	... original sentence is much easier to understand than the simplified one.
2	... original sentence is somewhat easier to understand than the simplified one.
3	... both sentences are equally easy/difficult to understand.
4	... simplified sentence is somewhat easier to understand than the original one.
5	... simplified sentence is much easier to understand than the original one.

Table 5: Definition of simplicity scores in expert evaluation.

Expert Evaluation

To better assess simplifications performed by different sentence simplification systems and their compliance with easy-to-read guidelines, we propose a novel expert evaluation and a detailed set of rules how to judge whether the transformation made by the system results in a simpler form (Table 2). For each language, we ask two expert annotators to perform the assessment. We provide them with the above-mentioned set of rules, and an online editing tool which highlights the differences between the original and simplified sentences.

The annotators are asked to count several types of lexical and syntactic transformations, and judge their correctness. The transformation is correct if it satisfies all three conditions: (1) preserves the original meaning; (2) is grammatical; and (3) results in a simpler phrase/sentence(s) according to the evaluation guidelines provided. If the conditions (1) and (3) are satisfied, but the transformation results in a small grammatical error (e.g. verb in plural instead of singular form), the transformation is semi-correct and receives a 0.5 score (instead of 1 for a completely correct transformation). The annotators are instructed to separately count and evaluate phrase level lexical transformations (everything beyond unigrams on either source or target side), sentence splitting, reordering within a clause, removal and addition of information. For each pair of original-simplified sentences, the annotators are requested to assign a meaning preservation score and simplicity score on a 1–5 scale (Tables 4 and 5).

The annotators are requested to compare their results, reach the consensus, and provide us with their final joint result. Several examples with correct and incorrect transformations are presented in Table 3. According to the guidelines for assigning simplicity and meaning preservation scores (Tables 4 and 5), the first sentence automatically simplified by MUSS-sup system in Table 3 would get the score 3 for meaning preservation (as it lost the essential information that *Graham graduated with BA in anthropology*) and score 5 for simplicity (due to sentence splitting, lexical simplification, and less information to process).

Results and Discussion

English Sentence Simplification

Standard Evaluation. We use SARI score to automatically compare our systems with previously proposed state-of-the-art sentence simplification systems with various architectures: the rule-based YATS system (Ferrés et al. 2016), phrase-based MT (Wubben, van den Bosch, and Krahmer 2012), encoder-decoder model (LSTM) with reinforcement learning Dress-LS (Zhang and Lapata 2017), original-transformer-based model DMASS+DCSS (Zhao et al. 2018), original transformer-based model with control tokens ACCESS (Martin et al. 2020), and transformer-based model (BART) with control tokens MUSS-sup (Martin et al. 2021). The ACCESS and MUSS-sup systems use the same

System	Lexical-all		Lexical-phrase		Reorder		Split		Remove		Add		Same	M	S
	All	Corr.	All	Corr.	All	Corr.	All	Corr.	All	Corr.	All	Corr.			
T5-base	**80**	**91%**	**49**	**90%**	**20**	**70%**	**17**	94%	**22**	**59%**	**16**	**87%**	**2%**	**4.2**	**4.3**
MUSS-sup	69	77%	34	82%	6	50%	16	**100%**	16	41%	7	71%	**2%**	4.1	3.8

Table 6: Results of the expert analysis for English, done on 50 randomly selected instances from MTurk test set, for two best performing systems (both systems were analysed for their output on the same 50 instances). The columns *Corr.* show the percentage of all cases of the respective category that were marked as correct. The column *Same* shows the percentage of sentences that were not changed by the system. Better scores in each category are presented in bold. Differences in M and S scores for the two systems are not statistically significant (Wilcoxon's sign rank test; $p < 0.01$).

System	Type	ASSET	MTurk
YATS	rule-based	34.4	37.4
PBMT-R	phrase-based MT	34.6	38.0
Dress-LS	LSTM+reinfor.	36.6	37.0
DMASS+DCSS	transformer	36.7	39.9
ACCESS	transf.+control	40.1	41.4
MUSS-sup	BART+control	**43.6**	**42.6**
mBART (our)	mBART+control	40.4	41.4
mT5-base (our)	mT5+control	42.0	41.2
T5-base (our)	T5+control	**44.9**	**43.3**

Table 7: SARI scores for English sentence simplification on two test sets (ASSET and MTurk), each with 359 instances. Higher scores indicate better outputs.

System	G	M	S
YATS	3.58*$_{\pm 0.14}$	3.54$_{\pm 0.14}$	3.25*$_{\pm 0.13}$
MUSS-sup	**3.99**$_{\pm 0.13}$	3.54$_{\pm 0.13}$	3.66$_{\pm 0.12}$
T5-base (our)	3.91$_{\pm 0.12}$	**3.58**$_{\pm 0.13}$	**3.68**$_{\pm 0.12}$

Table 8: Human evaluation scores (mean value with 95% confidence interval) for English on 50 randomly selected MTurk test instances. Higher scores indicate better outputs. Results marked with an '*' are significantly lower than the best ones (paired t-test; p<0.01).

four control tokens as our models (mBART, mT5-base, and T5-base), and the same training dataset. The only difference among those five systems (ACCESS, MUSS-sup, mBART, mT5-base, and T5-base) is the transformer model that is used. Our T5-base system achieves higher SARI score than all previously proposed systems on both test sets (Table 7). Overall, the results show superiority of transformer-based models with control tokens over all other approaches. We further perform human evaluation of grammaticality, meaning preservation, and simplicity, by five Amazon Mechanical Turk workers (all native English speakers) of several systems: our T5-base (as the best performing system), MUSS-sup (as the best performing previous system), and YATS (as the rule-based system). The results are presented in Table 8. The output of MUSS-sup and our T5-base are rated similarly. Both systems produce simpler and more grammatical sentences than YATS.

Expert evaluation. The results of the expert evaluation for English, performed on the same instances used for the crowdsourced human evaluation, are presented in Table 6. T5-base outperforms MUSS-sup by almost all metrics. The main issue found with MUSS-sup is the removal of essential parts that results in lower overall meaning preservation score (M). Two instances that illustrate those phenomena were presented earlier, in Table 3. The fewer number of lexical simplifications found in MUSS-sup and higher percentage of errors among those, led to a lower overall simplicity score (S). Among the additions made by MUSS-sup, only one was a hallucination: *"...on the steps of Michigan Union."* → *"...on the steps of Michigan Union University."*. The addition performed by MUSS-sup in one case led to a transformation of a sentence in present tense into a hypothetical sentence. All other additions made by MUSS-sup were correct. They were necessary to preserve grammaticality during reordering and sentence splitting. Among the additions made by T5-base, we found only one case of hallucination.

Overall, we found that both systems perform a range of distinct simplification operations. For each of the 16 rules from Table 2, we found at least one example of simplified sentence that is simpler than the original according to that rule in the output of each system. For example, we found two cases of passive to active voice conversion (e.g. *"Fives is a British sport believed to...* → *"Fives is a British sport. Many people think..."*) performed by T5-base, and one by MUSS-sup. All three were correct.

When interpreting the results in Table 6, it is important to remember that the only difference between the architectures used in T5-base and MUSS-sup is the transformer model (T5-base vs. BART). Both systems are trained with the same Wiki-Large dataset, and use the same four control tokens. Interestingly, we only found two instances for which both systems produced identical outputs.

Spanish Sentence Simplification

Standard Evaluation. For Spanish sentence simplification, we calculate SARI scores on the test set (350 instances) for the output of our three systems (mT5-base, mT5-large, and mBART), and the only two previously proposed fully-fledged systems: the rule-based system Simplext (Saggion et al. 2015), and the unsupervised MUSS system (Martin et al. 2021) which uses the combination of mBART with four control tokens (Table 10). The only difference between our mBART system and MUSS-unsup is that our system was trained with complex-simple sentence pairs from Spanish Newsela (7,414 sentence pairs), whereas the MUSS-unsup was trained with the web-mined paraphrases (996,609 sentence pairs). The mBART, mT5-large, and MUSS-unsup all

System	Lexical-all		Lexical-phrase		Reorder		Split		Remove		Add		Same	M	S
	All	Corr.	All	Corr.	All	Corr.	All	Corr.	All	Corr.	All	Corr.			
mT5-large	20	40%	5	40%	1	0%	**2**	**50%**	26	31%	4	0%	34%	3.4	**3.6**
mBART	40	45%	12	42%	0	NA	**2**	**50%**	22	57%	**6**	**83%**	24%	**3.6**	3.3
MUSS-unsup	**53**	**52%**	**27**	**57%**	33	24%	3	0%	**23**	**59%**	1	0%	**2%**	3.2	3.1

Table 9: Results of the expert analysis for Spanish, done on 50 randomly selected instances from the test set for three systems (the same 50 instances for all three systems). The columns *Corr.* show the percentage of all cases of the respective category that were marked as correct. The column *Same* shows the percentage of sentences that were not changed by the system. Better scores in each category are presented in bold. Differences in M and S are not significantly different (Wilcoxon's sign rank test; $p < 0.01$) for any pair of systems.

System	Type	SARI
Simplext	rule-based	33.5
MUSS-unsup	mBART+control	**36.8**
mT5-base (our)	mT5+control	32.7
mT5-large (our)	mT5+control	**36.9**
mBART (our)	mBART+control	**37.1**

Table 10: Results of Spanish sentence simplification.

System	G	M	S
MUSS-unsup	**4.52**$_{\pm 0.11}$	**3.96**$_{\pm 0.17}$	**3.51**$_{\pm 0.16}$
mT5-large (our)	4.43$_{\pm 0.13}$	3.81$_{\pm 0.18}$	3.19$_{\pm 0.18}$
mBART (our)	4.38$_{\pm 0.13}$	3.86$_{\pm 0.17}$	3.19$_{\pm 0.16}$

Table 11: Human evaluation scores (mean value with 95% confidence interval) for Spanish on 50 randomly selected test instances. Higher scores indicate better outputs. The differences in scores are not statistically significant (paired t-test; $p<0.01$) for any pair of systems.

achieve similar SARI scores, noticeably higher than those of the other two systems. Among them, the MUSS-unsup obtains the highest average scores for G, M, and S in the crowdsourced human evaluation (Table 11). However, the differences in G, M, and S scores between any pair of systems were not statistically significant.

Expert Evaluation. The results of the expert evaluation for Spanish are presented in Table 9. In comparison to the English T5-base system, the Spanish mT5-large system makes noticeably fewer lexical simplifications and sentence splittings, and has a higher percentage of erroneous ones. Both phenomena are very likely the result of much lower number of training instances for Spanish (7,414, as opposed to 296,402 for English) and the use of the multilingual instead of the monolingual transformer model.

According to the expert evaluation, MUSS-unsup performs more lexical simplifications than the other two Spanish sentence simplification models (especially mT5-large). However, those lexical transformations are found to be correct only in half of the cases (52%). The high percentage of errors made by MUSS-unsup resulted in noticeably lower average meaning preservation (M) and simplicity (S) scores. The most conservative system (mT5-large), which leaves 34% of the sentences unchanged, achieves the highest simplicity score among the three systems. Here is important to note that mBART and MUSS-unsup architectures differ only in the datasets they were trained with, their size and quality. MUSS-unsup was trained with a large number of web-mined paraphrases (996,609 sentence pairs), while mBART was trained with only 7,414 sentence pairs from a high quality Newsela dataset. These results indicate that, in transformer-based sentence simplification with these four control tokens, the size and quality of the training set strongly influence the number of transformations and their variety.

Conclusion

Automatic sentence simplification is envisioned to play a significant role in making everyday texts more accessible for wider populations thus ensuring their better social inclusion. In this study, we proposed several state-of-the-art sentence simplification systems for English and Spanish, using recently proposed transformer-based models coupled with a simplification control mechanism. We also proposed guidelines for expert human evaluation which takes into account recommendations for easy-to-read texts.

The extensive evaluation showed that proposed systems perform state-of-the-art sentence simplification in both English and Spanish, and that transformer-based systems with the chosen four-token control mechanism produce sentences that are simpler than the originals according to easy-to-read guidelines. All investigated transformer-based systems performed a wide range of simplification operations which lead to simpler output according to easy-to-read guidelines. In English sentence simplification, the results of expert evaluation indicate that the use of T5 leads to higher number of simplification operations and higher number of correct transformations than the use of mBART. In Spanish sentence simplification, the results of expert evaluation indicated that the size and the quality of the training data have influence on correctness of some transformations.

Acknowledgements

Our work is partly supported from the project Context-aware Multilingual Text Simplification (ConMuTeS) PID2019-109066GB-I00/AEI/10.13039/501100011033 awarded by Ministerio de Ciencia, Innovación y Universidades (MCIU) and by Agencia Estatal de Investigación (AEI) of Spain.

References

Akiba, T.; Sano, S.; Yanase, T.; Ohta, T.; and Koyama, M. 2019. Optuna: A Next-generation Hyperparameter Optimization Framework. *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*.

Aluísio, S. M.; Specia, L.; Pardo, T. A. S.; Maziero, E.; and De Mattos Fortes, R. P. 2008. Towards Brazilian Portuguese automatic text simplification systems. In *ACM Symposium on Document Engineering*, 240–248.

Alva-Manchego, F.; Martin, L.; Bordes, A.; Scarton, C.; Sagot, B.; and Specia, L. 2020. ASSET: A Dataset for Tuning and Evaluation of Sentence Simplification Models with Multiple Rewriting Transformations. In *Proceedings of ACL*, 4668–4679.

Alva-Manchego, F.; Martin, L.; Scarton, C.; and Specia, L. 2019. EASSE: Easier Automatic Sentence Simplification Evaluation. In *Proceedings of EMNLP-IJCNLP: System Demonstrations*, 49–54.

Alva-Manchego, F.; Scarton, C.; and Specia, L. 2020. Data-Driven Sentence Simplification: Survey and Benchmark. *Computational Linguistics*, 46(1): 135–187.

Carroll, J.; Minnen, G.; Canning, Y.; Devlin, S.; and Tait, J. 1998. Practical Simplification of English Newspaper Text to Assist Aphasic Readers. In *Proceedings of the AAAI'98 Workshop on Integrating AI and Assistive Technology*, 7–10.

Conneau, A.; Khandelwal, K.; Goyal, N.; Chaudhary, V.; Wenzek, G.; Guzmán, F.; Grave, E.; Ott, M.; Zettlemoyer, L.; and Stoyanov, V. 2020. Unsupervised Cross-lingual Representation Learning at Scale. In *Proceedings of ACL*, 8440–8451.

Cooper, M.; Reid, L. G.; Vanderheiden, G.; and Caldwell, B. 2010. Understanding WCAG 2.0. A guide to understanding and implementing Web Content Accessibility Guidelines 2.0. World Wide Web Consortium (W3C).

Coster, W.; and Kauchak, D. 2011. Learning to Simplify Sentences Using Wikipedia. In *Proceedings of the Workshop on Monolingual Text-To-Text Generation*, 1–9.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of NAACL-HLT, Volume 1 (Long and Short Papers)*, 4171–4186.

Ferrés, D.; Marimon, M.; Saggion, H.; and AbuRa'ed, A. 2016. YATS: Yet Another Text Simplifier. In *Proceedings of the 21st International Conference on Applications of Natural Language to Information Systems*, 335–342.

Flesch, R. 1949. *The Art of Readable Writing*. New York: Harper.

Freyhoff, G.; Hess, G.; Kerr, L.; Tronbacke, B.; and Van Der Veken, K. 1998. *Make it Simple, European Guidelines for the Production of Easy-toRead Information for People with Learning Disability*. ILSMH European Association, Brussels.

Horn, C.; Manduca, C.; and Kauchak, D. 2014. Learning a Lexical Simplifier Using Wikipedia. In *Proceedings of ACL*, ACL, 458–463.

Hwang, W.; Hajishirzi, H.; Ostendorf, M.; and Wu, W. 2015. Aligning Sentences from Standard Wikipedia to Simple Wikipedia. In *Proceedings of NAACL-HLT*, 211–217.

Kariuk, O.; and Karamshuk, D. 2020. CUT: Controllable Unsupervised Text Simplification. arXiv:2012.01936.

Karreman, J.; van der Geest, T.; and Buursink, E. 2007. Accessible Website Content Guidelines for Users with Intellectual Disabilities. *Journal of Applied Research in Intellectual Disabilities*, 20: 510–518.

Kintsch, W.; and van Dijk, T. A. 1978. Towards a model of text comprehension and production. *Psychological Review*, 85: 363–394.

Levenshtein, V. 1965. Binary codes capable of correcting deletions, insertions, and reversals. *Soviet physics. Doklady*, 10: 707–710.

Lewis, M.; Liu, Y.; Goyal, N.; Ghazvininejad, M.; Mohamed, A.; Levy, O.; Stoyanov, V.; and Zettlemoyer, L. 2020. BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. In *Proceedings of ACL*, 7871–7880.

Liu, Y.; Gu, J.; Goyal, N.; Li, X.; Edunov, S.; Ghazvininejad, M.; Lewis, M.; and Zettlemoyer, L. 2020. Multilingual Denoising Pre-training for Neural Machine Translation. *Transactions of ACL*, 8: 726–742.

Martin, L.; de la Clergerie, É.; Sagot, B.; and Bordes, A. 2020. Controllable Sentence Simplification. In *Proceedings of LREC*, 4689–4698.

Martin, L.; Fan, A.; Éric de la Clergerie; Bordes, A.; and Sagot, B. 2021. MUSS: Multilingual Unsupervised Sentence Simplification by Mining Paraphrases. arXiv:2005.00352.

Mencap. 2002. *Am I making myself clear? Mencap's guidelines for accessible writing*. Mencap, London.

Nisioi, S.; Štajner, S.; Ponzetto, S. P.; and Dinu, L. P. 2017. Exploring Neural Text Simplification Models. In *Proceedings of ACL*, 85–91.

Nomura, M.; Nielsen, G. S.; and Tronbacke, B. 1997. Guidelines for Easy-to-Read Materials. Technical report, FLA, Library Services to People with Special Needs Section.

OECD. 2013. OECD Skills Outlook 2013: First Results from the Survey of Adult Skills. Technical report, OECD Publishing.

Orăsan, C.; Evans, R.; and Mitkov, R. 2018. *Intelligent Text Processing to Help Readers with Autism*, 713–740. Cham: Springer International Publishing.

Papineni, K.; Roukos, S.; Ward, T.; and Zhu, W.-J. 2002. BLEU: A Method for Automatic Evaluation of Machine Translation. In *Proceedings of ACL*, 311–318.

Pavlick, E.; and Callison-Burch, C. 2016. Simple PPDB: A Paraphrase Database for Simplification. In *Proceedings of ACL (Volume 2: Short Papers)*, 143–148. ACL.

PlainLanguage. 2011. Federal Plain Language Guidelines.

Raffel, C.; Shazeer, N.; Roberts, A.; Lee, K.; Narang, S.; Matena, M.; Zhou, Y.; Li, W.; and Liu, P. J. 2020. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *Journal of Machine Learning Research*, 21(140): 1–67.

Saggion, H. 2017. *Automatic text simplification*. Morgan & Claypool Publishers.

Saggion, H.; Štajner, S.; Bott, S.; Mille, S.; Rello, L.; and Drndarević, B. 2015. Making It Simplext: Implementation and Evaluation of a Text Simplification System for Spanish. *ACM Transactions on Accessible Computing (TACCESS)*, 6(4): 14.

Scarton, C.; and Specia, L. 2018. Learning Simplifications for Specific Target Audiences. In *Proceedings of ACL (Volume 2: Short Papers)*, 712–718.

Sheang, K. C.; and Saggion, H. 2021. Controllable Sentence Simplification with a Unified Text-to-Text Transfer Transformer. In *Proceedings of INLG*, 341–352.

Siddharthan, A. 2006. Syntactic Simplification and Text Cohesion. *Research on Language and Computation*, 4(1): 77–109.

Siddharthan, A.; and Mandya, A. 2014. Hybrid text simplification using synchronous dependency grammars with hand-written and automatically harvested rules. In *Proceedings of EACL*, 722–731.

Specia, L. 2010. Translating from Complex to Simplified Sentences. In *Proceedings of PROPOR*, 30–39.

Štajner, S.; and Hulpuş, I. 2018. Automatic Assessment of Conceptual Text Complexity Using Knowledge Graphs. In *Proceedings of COLING*, 318–330.

Štajner, S.; and Saggion, H. 2018. Data-Driven Text Simplification. In *Proceedings of COLING: Tutorial Abstracts*, 19–23.

Sulem, E.; Abend, O.; and Rappoport, A. 2018. BLEU is Not Suitable for the Evaluation of Text Simplification. In *Proceedings of EMNLP*, 738–744.

Tanprasert, T.; and Kauchak, D. 2021. Flesch-Kincaid is Not a Text Simplification Evaluation Metric. In *Proceedings of the 1st Workshop on Natural Language Generation, Evaluation, and Metrics (GEM 2021)*.

Vásquez-Rodríguez, L.; Shardlow, M.; Przybyła, P.; and Ananiadou, S. 2021. Investigating Text Simplification Evaluation. In *Findings of ACL*.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. In *Advances in neural information processing systems*, 5998–6008.

Štajner, S. 2021. Automatic Text Simplification for Social Good: Progress and Challenges. In *Findings of ACL*, 2637–2652.

Štajner, S.; Franco-Salvador, M.; Ponzetto, S. P.; Rosso, P.; and Stuckenschmidt, H. 2017. Sentence Alignment Methods for Improving Text Simplification Systems. In *Proceedings of ACL*, 97–102.

Štajner, S.; Franco-Salvador, M.; Rosso, P.; and Ponzetto, S. P. 2018. CATS: A Tool for Customized Alignment of Text Simplification Corpora. In *Proceedings of LREC*, 3895–3903.

Štajner, S.; Mitkov, R.; and Saggion, H. 2014. One Step Closer to Automatic Evaluation of Text Simplification Systems. In *Proceedings of the 3rd Workshop on Predicting and Improving Text Readability for Target Reader Populations (PITR) at EACL*, 1–10.

Štajner, S.; and Nisioi, S. 2018. A Detailed Evaluation of Neural Sequence-to-Sequence Models for In-domain and Cross-domain Text Simplification. In *Proceedings of LREC*, 3026–3033.

W3C. 2008. *Web Content Accessibility Guidelines (WCAG) 2.0*.

Wolf, T.; Debut, L.; Sanh, V.; Chaumond, J.; Delangue, C.; Moi, A.; Cistac, P.; Rault, T.; Louf, R.; Funtowicz, M.; Davison, J.; Shleifer, S.; von Platen, P.; Ma, C.; Jernite, Y.; Plu, J.; Xu, C.; Le Scao, T.; Gugger, S.; Drame, M.; Lhoest, Q.; and Rush, A. 2020. Transformers: State-of-the-Art Natural Language Processing. In *Proceedings of EMNLP: System Demonstrations*, 38–45.

Wubben, S.; van den Bosch, A.; and Krahmer, E. 2012. Sentence Simplification by Monolingual Machine Translation. In *Proceedings of ACL: Long Papers - Volume 1*, 1015–1024.

Xu, W.; Callison-Burch, C.; and Napoles, C. 2015. Problems in Current Text Simplification Research: New Data Can Help. *Transactions of ACL*, 3: 283–297.

Xu, W.; Napoles, C.; Pavlick, E.; Chen, Q.; and Callison-Burch, C. 2016. Optimizing Statistical Machine Translation for Text Simplification. *Transactions of ACL*, 4: 401–415.

Xue, L.; Constant, N.; Roberts, A.; Kale, M.; Al-Rfou, R.; Siddhant, A.; Barua, A.; and Raffel, C. 2021. mT5: A Massively Multilingual Pre-trained Text-to-Text Transformer. In *Proceedings of NAACL-HLT*, 483–498.

Zhang, X.; and Lapata, M. 2017. Sentence Simplification with Deep Reinforcement Learning. In *Proceedings of EMNLP*, 584–594.

Zhao, S.; Meng, R.; He, D.; Saptono, A.; and Parmanto, B. 2018. Integrating Transformer and Paraphrase Rules for Sentence Simplification. In *Proceedings of EMNLP*, 3164–3173.

Zhu, Z.; Bernhard, D.; and Gurevych, I. 2010. A Monolingual Tree-based Translation Model for Sentence Simplification. In *Proceedings of COLING*, 1353–1361.