A Detailed Evaluation of Neural Sequence-to-Sequence Models for In-domain and Cross-domain Text Simplification

Sanja Štajner[1,*], **Sergiu Nisioi**[2,*]
[1] DWS Research Group, University of Mannheim, Germany
[2] Human Language Technologies Research Center, University of Bucharest, Romania
sanja@informatik.uni-mannheim.de, sergiu.nisioi@fmi.unibuc.ro

Abstract

We present a detailed evaluation and analysis of neural sequence-to-sequence models for text simplification on two distinct datasets: Wikipedia and Newsela. We employ both human and automatic evaluation to investigate the capacity of neural models to generalize across corpora, and we highlight challenges that these models face when tested on a different genre. Furthermore, we establish a strong baseline on the Newsela dataset and show that a simple neural architecture can be efficiently used for in-domain and cross-domain text simplification.

Keywords: neural text simplification, sequence-to-sequence models, evaluation

1. Introduction

The aim of text simplification (TS) is to transform given texts into their syntactically and/or lexically simpler variants which are more understandable for the target population (e.g. children, non-native speakers, people with low literacy levels, or people with various kinds of cognitive or reading impairments). It is usually applied on the sentence level and encompasses three major simplification operations: sentence splitting, deletions, and lexical paraphrases (Xu et al., 2016).

In recent years, the problem of automated sentence simplification has often been addressed as the monolingual machine translation (MT) task of translating from original to simple sentences. The MT models used were, however, adapted to the specificities of the TS task, e.g. by adding phrasal deletions to the standard phrase-based MT model (Coster and Kauchak, 2011) to account for a common sentence shortening and phrasal deletions in TS, or by reranking the output of the phrase-based MT model (Wubben et al., 2012), since in the standard phrase-based MT model applied on TS, the first hypothesis tends to leave the input unchanged (Specia, 2010; Coster and Kauchak, 2011). Until recently, the state of the art for automated text simplification was the syntax-based machine translation system (SBMT) with specific optimizations for TS (Xu et al., 2016), such as the use of a large paraphrase database (PPDB) to boost the coverage of the phrasal simplifications, and the use of SARI, a tuning metric that particularly rewards simplicity (Xu et al., 2016).

Following more recent advancements in machine translation using neural networks, we proposed a neural text simplification system, which significantly outperformed the state of the art on various evaluation metrics for the Wikipedia dataset (Nisioi et al., 2017). Our model was constructed as a vanilla encoder-decoder architecture with global attention and input feeding. More recently, a neural network model fine-tuned using reinforcement learning (Zhang and Lapata, 2017) has been proposed for text simplification, the authors reporting several improvements over the previous systems. Given that they were proposed around the same time, the two neural TS models have not yet been directly compared.

One commonly raised issue with most supervised TS systems is the usage of English Wikipedia – Simple English Wikipedia (EW–SEW) sentence-aligned corpora for training the systems, especially since the quality of SEW for modeling TS has often been disputed (Amancio and Specia, 2014; Štajner et al., 2015; Xu et al., 2016). Another parallel corpus of original news articles and their manually simplified versions on four different complexity levels, following strict guidelines and quality control – the Newsela corpus (Newsela, 2016)[1] – has been recently released. Xu *et al.* (2016) show that it has better potential than the EW–SEW dataset for the TS task. However, up until recently, the Newsela corpus was only provided with alignments at the document level.

We use the freely available software[2] for sentence-alignment across different Newsela levels (Štajner et al., 2017; Štajner et al., 2018), and then train neural text simplifications models on the sentence-aligned Newsela and EW–SEW (Hwang et al., 2015) datasets. We compare our systems with the SBMT system (Xu et al., 2016) and the recently proposed state-of-the-art reinforcement learning NTS model (Zhang and Lapata, 2017) to show that a simple neural architecture can be efficiently used for in-domain and cross-domain TS.

Last but not least, we provide a detailed human and automatic evaluation of neural sequence-to-sequence models trained and tested in-domain and cross-domain on each of the two corpora, to discuss the ability of these models to generalize across registers.

2. Methodology

In this section, we describe our neural text simplification (NTS) models, the datasets used for training and testing, and the evaluation procedures.

[*] Both authors have contributed equally to this work

[1] Freely available for research upon request at `Newsela.com`

[2] `https://github.com/neosyon/SimpTextAlign`

2.1. Models

Following the success of neural sequence-to-sequence models in TS (Nisioi et al., 2017), our simplification systems are based on neural networks (Graves, 2012) with global attention in combination with input feeding (Luong et al., 2015). We use the *OpenNMT* framework (Klein et al., 2017) to train and build an architecture with two LSTM layers (Hochreiter and Schmidhuber, 1997), 500 hidden units, embedding size of 300, and 0.3 dropout probability (Srivastava et al., 2014). We train the model for 14 epochs, regardless of the dataset used, with stochastic gradient descent optimizer and a learning rate decay of 0.7 starting from epoch 7. To be able to have comparable results in-domain and across multiple corpora, we do not use pre-trained embeddings. Several changes made in the meantime for the *OpenNMT* framework and other third party libraries trigger different results than the ones reported by Nisioi et al. (2017). To be able to have a comparable overview across systems, we set the beam size to 12 and re-generate the output of our systems. The models trained on Wikipedia together with the outputs on the same corpus are publicly released.[3]

It may be the case that this type of sequence-to-sequence architecture has by now become a standardized vanilla model for machine translation (Bojar et al., 2017). Nevertheless, for text simplification, we notice a few particularities that helped improve the learning. First of all, for our datasets, we do not use sub-word models (Sennrich et al., 2015; Luong and Manning, 2016), since English does not present rich morphological agglutinations. Secondly, we observe that a size of 300 for internal word embeddings is enough for both Newsela and Wikipedia datasets, the system producing lexical changes without the use of any external information. And lastly, we note the importance of keeping a reduced size of the vocabulary - no more than 50,000 words. On the one hand, this limits the amount of low frequency words that the model learns in order to produce lexical simplifications, but on the other hand, it ensures the grammaticality and meaning preservation of the output, by simply re-using words from the source sentence.

For simplifying sentences, we use beam search to sample four hypotheses corresponding to the given input. As shown in our previous work (Nisioi et al., 2017), the first hypothesis generated by a sequence-to-sequence model is not always the most relevant for text simplification. The reason behind this is that the model, by default, tends to preserve the meaning and the words from the input, whereas, the hypotheses with lower likelihood scores tend to present a greater degree of content reduction and lexical changes. This is further corroborated by the human and automatic evaluation presented in Section 3. More explicitly, we model the hypothesis number as a hyper-parameter that we select after the model finished training. For each model trained on Wikipedia and on Newsela, we generate predictions on each individual development set, obtaining 4 hypotheses corresponding to the training-test data pairs. When predicting on the test set, we choose the hypothesis based on the maximum average SARI score of that hypothesis number on the development set, as shown in Table 1.

	Train		
	Newsela	**Wikipedia**	
Dev set	SARI score		**Hyp.**
Newsela	29.71	29.55	1
	38.89	33.40	2
	39.34	33.48	3
	39.25	**33.79**	4
Wikipedia	30.99	31.19	1
	35.10	**36.39**	2
	35.81	35.71	3
	35.54	36.23	4

Table 1: SARI scores on Newsela and Wikipedia predictions on the development set by models trained on the two datasets.

2.2. Datasets

2.2.1. Newsela Datasets

Newsela offers original news articles and their manual simplifications at four different levels of simplification. We automatically sentence-aligned the English part of the Newsela corpus across different levels using the freely available, recently released CATS software particularly made for this purpose (Štajner et al., 2017; Štajner et al., 2018). A thorough human evaluation performed on over 3,000 sentence pairs showed that the accuracy of automatic alignment between the sentences from two neighboring Newsela levels of simplification is between 83% (for levels 3 and 4) and 98% (for levels 0 and 1), while the alignment method aligns sentences from the hardest and the easiest levels (0 and 4), with only 58% accuracy (Štajner et al., 2017). Therefore, we aligned and used only the neighboring Newsela levels (0–1, 1–2, 2–3, and 3–4) as training data. The alignment procedure has been shown to successfully align sentences with full semantic overlap, which can be used to model lexical and syntactic paraphrases, as well as those with only partial semantic overlap, which can be used to model deletion and addition operations (Štajner et al., 2018). The CATS software also provides '1-n' alignments, which can be used to model sentence splitting, where appropriate (Štajner et al., 2018).

The Newsela corpus is organized in unique files based on the topics being addressed, covering approximately 2,000 topics. We split our training, development, and test data disjointly based on the topic files, ensuring that the sentences from the same story (regardless of their complexity levels) never appear in both the training and test data. The exact size of each subset is provided in Table 2. For testing purposes, we use the common sentences from the alignments between 0–1, 0–2, 0–3, and 0–4 to create multiple simplification hypotheses for more accurate calculation of automatic evaluation scores. Both the development references and the test references used to report the scores in Tables 1 and 4 are created from these types of alignments.

[3] https://github.com/senisioi/NeuralTextSimplification

	#topic files	#sentences	#tokens O	#tokens S
train	1,835	298,814	5,791,417	5,823,546
dev	56	9,372	180,682	181,742
test	19	655	17,132	17,506

Table 2: The size in terms of topic files, sentences, and tokens for the original (O) and simplified (S) versions of the dataset that we used for training, testing, and developing our Newsela models.

	#sentences	#tokens O	#tokens S
train	284,677	7,400,499	5,634,834
dev	16,000	349,944	308,856
test	359	8,110	7,957

Table 3: Statistics regarding the number of sentences and tokens for the original (O) and simplified (S) versions of Wikipedia dataset used to train our models.

2.2.2. Wikipedia Datasets

Our Wikipedia dataset consists of the latest sentence-aligned version (Hwang et al., 2015) based on manual and automatic alignments between standard English Wikipedia and Simple English Wikipedia (EW–SEW). We discard the uncategorized matches, and use only *good matches* and *partial matches* which were above the 0.45 threshold (Hwang et al., 2015), totaling to 280K aligned sentences (around 150K full matches and 130K partial matches). Unlike the earlier EW–SEW version[4] (Kauchak, 2013) which only contains *full matches* (167K pairs), the newer dataset that we use also contains *partial matches*, and is thus not only larger, but also allows for learning sentence shortening transformations. From this dataset, we remove those sentence pairs whose original sentences are present in the Wikipedia test set compiled by Xu et al. (2016). We also opt for this test set, as it contains, for each of the 359 original sentences, eight manually simplified versions that can be used as multiple references for more accurate calculation of automatic evaluation scores. Statistics regarding the size of the Wikipedia datasets are rendered in Table 3. Unlike the Newsela datasets, the Wikipedia datasets do not contain examples of sentence splitting, as the original EW–SEW dataset (Hwang et al., 2015) only contains one-to-one sentence alignments.

2.3. Evaluation Procedures

2.3.1. Automatic Evaluation

BLEU (Papineni et al., 2002) is a standardized metric for machine translation evaluation that reports a similarity score between the output of a system and the 'gold standard' references. In this paper, we report BLEU with NIST smoothing as implemented in NLTK (Bird et al., 2009). One downside of this score for text simplification, however, stems from the sole comparison of the output against references without considering the initial sentence. Based on this idea, a metric that compares system output against references and against input - SARI (Xu et al., 2016), has been proposed to reward additions, copying, and deletions from the input that are present in the output and references. Unlike BLEU, the SARI score has been shown to better predict the simplicity of the output (Xu et al., 2016).

To account for all three aspects (grammaticality, meaning preservation, and simplicity) of the output sentences, we report both BLEU and SARI scores. Nevertheless, both those automatic metrics are used only as additional evaluation metrics, while the discussion is based solely on a detailed human evaluation.

2.3.2. Human Evaluation

For human evaluation, we follow the procedure proposed in our recent paper (Nisioi et al., 2017) and report:

- The percentage of sentences which undergone at least one change;
- The total number of changes;
- The percentage of correct changes;
- Grammaticality of the simplified sentence;
- Meaning preservation of the simplified sentence;
- Relative simplicity of the simplified sentence in comparison to the original sentence.

The total number of changes counts the number of changes regardless of their type (lexical changes, syntactic changes such as phrase reordering or sentence splitting, deletions, and additions). In the case of phrasal substitutions, the changes of a whole phrase (e.g. "*become defunct*" → "*was dissolved*") are counted as one change. In the case of content reduction (deletion), we instructed the annotators to count the deletion of each array of consecutive words as one change. The count was performed by two native English speakers. The sentences for which the two annotators did not agree were given to a third annotator to obtain the majority vote.

Those changes that preserve the original meaning and grammaticality of the sentence (assessed by two native English speakers) and, at the same time, make the sentence easier to understand (assessed by two non-native fluent English speakers) are marked as *Correct*. Given that our systems were trained to model not only full lexical and syntactic paraphrasing, but also content reduction (due to the partial matches in our training datasets), we instructed the annotators to consider the meaning unchanged if the main information of the sentence was retained and unchanged. The sentences for which the two annotators did not agree were given to a third annotator to obtain the majority vote.

Grammaticality (*G*) and meaning preservation (*M*) of the simplified sentences were assessed by three native English speakers using a 1–5 Likert scale (1 → very bad; 5 → very good). The final scores were computed as the arithmetic mean of the scores by the three annotators. Only those sentences which have undergone at least one modification are taken into account for calculating the G and M scores. This way, we make sure that the systems which leave many input sentences unchanged do not get rewarded for that and result in higher G and M scores, as the sentences which are left unchanged always get the highest G and M score (they are

[4] `http://www.cs.pomona.edu/~dkauchak/simplification/`

grammatically correct and have exactly the same meaning as the original sentence).

The simplicity (*S*) score was assigned by three non-native fluent English speakers. The annotators were shown original (reference) sentences and target (output) sentences, one pair at the time, and asked whether the target sentence is: +2 → much simpler; +1 → somewhat simpler; 0 → equally difficult; -1 → somewhat more difficult; -2 → much more difficult, than the reference sentence. The final simplicity score was calculated as the arithmetic mean of the scores assigned by all three annotators.

We did not explicitly instruct the annotators regarding the influence of grammaticality on the simplicity score. The post-evaluation analysis revealed that ungrammatical and meaningless sentences were penalized by receiving a negative simplicity score by all three annotators. Therefore, if one was to chose just one evaluation measure to compare the performances of different TS systems, the simplicity score assigned in this manner would probably be the right choice. However, as the annotators were not instructed to take into account meaning preservation while assigning the simplicity score, the meaning preservation scores of the systems would have to be additionally checked.

Here is also important to note that while the correctness of changes takes into account the influence of each individual change on the grammaticality, meaning preservation and simplicity of a sentence, the G, M, and S scores take into account the mutual influence of all changes within a sentence.

3. Results and Discussion

We first explore the BLEU and SARI scores on the first four hypothesis in both in-domain and cross-domain scenarios (Section 3.1). Next, we evaluate (automatically and manually) our NTS models for in-domain and cross-domain text simplification (Section 3.2), comparing the models which always choose the default first hypothesis (as the baselines of our models) with those that use the SARI score to choose the best hypothesis. Finally, we compare the performances of our best models with the performances of the state-of-the-art SBMT model (Xu et al., 2016) and the state-of-the-art reinforcement learning NTS model (Zhang and Lapata, 2017) on the Wiki test set (Xu et al., 2016) in Section 3.3.

3.1. Automatic Evaluation of Hypotheses

We automatically evaluate models trained on each of our two datasets, in a pairwise fashion, against each test set. Table 4 contains the results of the automatic evaluation for each beam search hypothesis from 1 to 4 (last column).

	Train				
	Newsela		**Wikipedia**		
Test	**BLEU**	**SARI**	**BLEU**	**SARI**	**Hyp.**
Newsela	**77.06**	28.21	**64.16**	30.81	1
	71.66	37.06	59.14	33.69	2
	70.51	**38.84**	57.81	33.67	3
	70.31	37.76	58.43	**34.0**	4
Wikipedia	**89.49**	30.33	**84.69**	30.54	1
	84.75	35.00	77.57	**35.78**	2
	83.8	**36.48**	77.21	35.67	3
	83.57	36.15	75.77	35.76	4

Table 4: SARI and BLEU scores on Newsela and Wikipedia predictions by models trained on the two datasets.

On the one hand, if we focus on the BLEU evaluation score, we notice that the first hypothesis, the one most likely given the beam search score, always obtains the highest BLEU score, regardless of the training-test pairs. On the other hand, if we focus on the TS-specific metric SARI (Xu et al., 2016), the best scores are never obtained by the first hypotheses, but rather by the ones with lower probability and less content from the input. This is expected given that SARI especially rewards the output which is the least similar to the input. For example, the following sentence:

- *In its pure form, dextromethorphan occurs as a white powder.*

as an input to the model trained on the Wikipedia dataset, generates the following four hypotheses:

1. *In its pure form, dextromethorphan occurs as a white powder.*
2. *Dextromethorphan occurs as a white powder.*
3. *Dextromethorphan is a white powder.*
4. *It is a white powder.*

For this particular short sentence, the first hypothesis is identical to the input, whereas the lower likelihood hypotheses 2, 3 and 4 present more traits of simplification.

3.2. Sequence-to-Sequence Models

The evaluation scores for our in-domain and cross-domain NTS systems, both with default ranking of the hypotheses and with reranking of the hypotheses according to the SARI score on the dev set are reported in Table 5.

3.2.1. Reranking

We can notice that reranking of the output according to the SARI metric (instead of using the default first hypothesis h1) always leads to a significant increase in percentage of sentences that were changed (up to more than three times more sentences changed in the case of in-domain TS on the Newsela dataset), and higher grammaticality (G) and meaning preservation (M) scores. In most of the cases, it also leads to a substantial increase in simplicity score (S) and in percentage of correct changes. The most striking difference between the system that chooses the default hypothesis h1 and the one that chooses the hypothesis with the best SARI scores on the dev set is achieved in the case of in-domain TS on the Newsela dataset. One potential reason for this might be that our Newsela training data contains only the consecutive alignments (e.g. 0–1, 1–2, 2–3, and 3–4) making the model learn small changes that appear from one level to another.

The only case in which the system with the default hypotheses h1 outperforms the system with the reranked output is the cross-domain TS where the systems were trained on the Wikipedia dataset and tested on the Newsela dataset. In

Domain	Training-Test	Rerank.	Hypothesis	changed	Changes		Scores		S	SARI	BLEU
				sent.	Total	Correct	G	M			
In	News - News	default	h1	27.1%	23	21.7%	4.52	2.31	+0.02	28.21	**77.09**
In	News - News	SARI	h3	**90.0%**	**76**	**54.1%**	**4.97**	**3.87**	+**0.50**	**38.84**	70.51
In	Wiki - Wiki	default	h1	41.4%	37	48.6%	4.59	3.41	+0.30	30.54	**84.69**
In	Wiki - Wiki	SARI	h2	**87.1%**	**78**	**59.0%**	**4.77**	**4.05**	+**0.49**	**35.78**	77.57
Cross	Wiki - News	default	h1	47.1%	46	**28.3%**	3.87	2.55	+**0.23**	30.81	**64.16**
Cross	Wiki - News	SARI	h4	**85.5%**	**77**	25.0%	**4.48**	**3.52**	+0.21	**34.00**	58.43
Cross	News - Wiki	default	h1	40.0%	37	18.9%	3.86	2.90	+0.04	30.33	**89.49**
Cross	News - Wiki	SARI	h3	**97.1%**	**102**	**23.4%**	**4.34**	**3.19**	+**0.28**	36.48	83.80

Table 5: Human evaluation results on the first 70 test sentences (the highest scores obtained on each test set by each evaluation criterion are shown in bold) and automatic evaluation measures (BLEU and SARI) on the full test sets.

Approach	changed	Changes		Scores		S	SARI	BLEU
	sent.	Total	Correct	G	M			
Our NTS - training on Wiki - h2 (best SARI)	87.1%	78	**59.0%**	**4.77**	**4.05**	+**0.49**	35.78	77.57
Our NTS - training on News - h3 (best SARI)	**97.1%**	102	23.4%	4.34	3.19	+0.28	36.48	**83.80**
SBMT (SARI+PPDB) (Xu et al., 2016)	82.9%	**143**	34.3%	4.28	3.57	+0.03	**38.59**	73.62
Dress-LS (Zhang and Lapata, 2017)	67.1%	63	42.9%	4.27	3.80	+0.14	32.74	81.16

Table 6: Human evaluation results on the first 70 test sentences (the highest scores obtained on each test set by each evaluation criterion are shown in bold) and automatic evaluation measures (BLEU and SARI) on the full test sets of our two best systems (in-domain and cross-domain) and the state-of-the-art systems on the Wiki test set.

this case, the default hypotheses led to a slightly higher percentage of correct changes and a slightly higher simplicity gain (S). Nevertheless, the achieved grammaticality (G) and meaning preservation (M) scores were substantially lower for the system with the default hypothesis h1.

The examples *1a–1d* and *2a–2c* in Table 8 illustrate the cases in which the default hypotheses h1 fails to make any changes to the input, while the best SARI ranked hypotheses improve the simplicity of the sentence by introducing correct changes.

3.2.2. In-domain vs. Cross-domain Simplification

When we compare the performances of our NTS systems on in-domain and cross-domain tasks, we notice that, in both cases, they result in similar percentages of changed sentences. However, the cross-domain TS results in substantially lower percentage of correct changes and substantially lower simplicity gain (S). The cross-domain NTS also achieves lower grammaticality (G) and meaning preservation (M) scores. Two examples of ungrammatical output obtained by the cross-domain NTS, and one example of wrong lexical substitution applied in cross-domain NTS are shown in Table 8 (examples *3d–3e*, and *4d* for the ungrammatical output, and example *2c* for the wrong lexical substitution). In all three cases, the in-domain NTS resulted in grammatical output.

3.2.3. Wikipedia vs. Newsela Datasets

The systems with the default hypotheses h1 achieve higher evaluation scores for the Wikipedia than for the Newsela dataset. We can only explain this phenomenon by the amount of variety from the two training sets: the Wikipedia corpus contains a high variety of topics that can appear in both training and test datasets, while the Newsela training set contains a fixed amount of stories that are repeated for different levels of simplification. For example, the training data may contain exactly the same sentence both as original and simplified example. What for levels 0–1 can be an example of simplification, for levels 1–2 it is a complex sentence that needs to be further simplified, and so on. In this case, the default hypothesis will likely be biased. Therefore, we believe it is essential to sample different hypotheses from the model in order to get multiple sources of truth. If the hypotheses with the best SARI scores on the dev set are used instead, then our NTS systems perform equally well on both domains.

3.3. Comparison with the State of the Art

Given that the SBMT system (Xu et al., 2016) and all its components are not freely available, and the Newsela splits used for the Dress-LS systems (Zhang and Lapata, 2017) were not available at the time of our experiments, we were able to directly compare our systems with those two state-of-the-art systems only on the Wiki test set (Xu et al., 2016), for which the outputs of both those systems are freely available.

The results of the manual evaluation (Table 6) show that on the Wiki test set, our in-domain NTS model (trained on the Wikipedia dataset) outperforms both state-of-the-art TS systems (SBMT and Dress-LS). It results in higher percentage of sentences which undergone at least one change, higher percentage of correct changes, and higher grammaticality, meaning preservation and simplicity scores. More importantly, the difference in the obtained simplicity gain (S) is substantial (+0.49 as opposed to +0.03 and +0.14, respectively), as well as the difference in grammaticality and meaning preservation scores.

Train-Test	Split	Short	NE
Wiki–Wiki	0	24	2
Wiki–News	0	18	2
News–News	3	10	20
News–Wiki	5	15	3

Table 7: The number of sentence splittings, sentence shortenings (removal of at least five consecutive words, or an apposition, or a part of a noun phrase), and wrong NE replacements in the first 70 test sentences.

Interestingly, even our cross-domain NTS system (with the reranking of hypotheses according to the SARI scores on the dev set) performs better than the state-of-the-art TS systems. It outperforms both systems (SBMT and Dress-LS) by the number of sentences which undergone at least one change (97.1% instead of 82.9% and 67.1%, respectively) and by the simplicity gain (+0.28 as opposed to +0.03 and +0.14, respectively), while achieving similar grammaticality of the output (4.34 as opposed to 4.28 and 4.27, respectively). Although it has a lower percentage of correct individual changes, the output of our NTS model trained on Newsela achieves much higher simplicity gain (S) than the SBMT and Dress-LS systems, indicating thus that those fewer correct changes still have significant impact on the simplicity of the output. This is probably due to the sentence shortenings and sentence splittings learned by our NTS model (see Table 7).

For illustration, Table 8 contains two examples in which our NTS systems perform better than the SBMT and Dress-LS systems (examples *1a–1d* and *6a–6e*). The examples *3b* and *4b* show the wrong lexical substitutions performed by the SBMT system which led to low grammaticality and meaning preservation scores for that system. In example *5*, the Dress-LS system performed one correct lexical substitution which our NTS systems did not, but at the same time, due to the sentence shortening, an important piece of information was lost.

4. Further Analysis

To better understand some phenomena noticed during the manual evaluation, which are specific for our NTS approach and datasets, we count the number of sentence splittings, sentence shortenings and wrong named entity (NE) replacements on all our train-test combinations (Table 7).

The NTS models trained on the Newsela dataset were able to learn sentence splitting operations, unlike the models trained on the Wikipedia dataset. The second is not surprising, given that the Wikipedia dataset (Hwang et al., 2015) does not allow for one-to-many sentence alignments and therefore does not contain sentence splitting examples. However, when trained on a dataset that contains examples of sentence splittings (the Newsela dataset), our NTS models were able to learn this simplification operation and successfully apply it on a test set from either the same domain (examples *2a–2c*, Table 8) or from another domain (examples *5a–5b*, Table 8).

All our NTS models were able to successfully apply sentence shortening in all train-test combinations. The models trained on the Wikipedia dataset performed more sentence shortenings, which is probably due to the fact that the Wikipedia dataset contains an abundant amount of partial matches (which are good training material for sentence shortening), which is not the case in the Newsela training dataset which consists only of the sentence pairs from the neighboring levels.

The number of sentence splittings and sentence shortenings obviously reflects the type of training data and the way it was collected. Nevertheless, it is important that our NTS models seem to be able to learn whichever type of sentence transformation is present in the training dataset and apply it even on another domain and text genre.

The high number of NE errors found in the NTS models trained and tested on the Newsela dataset probably reflects the facts that: (1) news contain an abundant amount of named entities; and (2) we did not allow for the same topics/news stories in the training and test Newsela datasets, thus creating the ideal opportunity for the unseen NEs in the test set. As the Wikipedia dataset does not have the information about the exact article from which the sentence pair was taken, we could not use the same constraints on the Wikipedia datasets. That is probably the reason for the much lower number of NE errors by the NTS systems trained and tested on the Wikipedia dataset.

However, the large number of NE errors made by the NTS models trained and tested on the Newsela dataset does not seem to have greatly influenced the overall performance of the NTS systems (see Table 6). We believe this is due to the fact that wrong entity substitutions in a sentence do not damage its grammaticality as much as wrong substitutions of other words (see examples *1c*, *3b*, *4b*, and *6e* in Table 8). Since one wrong entity replacement does not make a sentence more difficult to understand (rather it changes its meaning) and sentence splitting significantly simplifies a sentence, the NTS model trained on the Newsela dataset was better ranked than the SBMT and Dress-LS systems for its relative simplicity to the original (the S score). The wrong substitutions clearly affect the meaning, and this is reflected in the results presented in Table 6, where our NTS system trained on the Newsela dataset obtained a lower M score than the SBMT and Dress-LS systems.

5. Conclusions

In recent years, text simplification was often modeled as the monolingual machine translation task of translating original sentences into their simpler variants. Following the latest trends in machine translation, several text simplification models with neural architecture were proposed last year and they became the state of the art in English TS.

In this study, we focused on a sequence-to-sequence model to investigate its performance for in-domain and cross-domain text simplification, providing detailed automatic and human evaluations. As expected, the in-domain NTS models achieve better SARI scores and the human evaluation scores compared to those of the cross-domain NTS models.

According to our results, neural networks prove once again to be powerful tools to model text simplification, presenting significant improvements over the earlier proposed syntax-

Ex.	System	Output
1a	Original-W, NTS-h1-any	She remained in the United States until 1927 when she and her husband returned to France.
1b	NTS-SARI-any	She **stayed** in the United States until 1927 when she and her husband returned to France.
1c	SBMT	She *is still* in the United States until 1927 when she and her husband returned to France.
1d	Dress-LS	She stayed in the United States until 1927 when she *was married* to France.
2a	Original-N, NTS-h1-any	Both newcomers and advanced learners trained together, but those with more experience were given more challenging training.
2b	NTS-SARI-in	Both newcomers and advanced learners trained together**. However,** those with more experience were given more challenging training.
2c	NTS-SARI-cross	Both newcomers and advanced *atheists* trained together, but those with more experience were given more challenging training.
3a	Original-W, NTS-h1-cross	Disney received a full-size Oscar statuette and seven miniature ones, presented to him by 10-year-old child actress Shirley Temple.
3b	SBMT	Disney won a full-size Oscar statue and seven *the mini*, *made* to him by *10 year child* actress Shirley Temple.
3c	NTS-h1-in	*It was* presented to him by 10-year-old child actress Shirley Temple.
3d	NTS-SARI-in, Dress-LS	Disney received a full-size Oscar statuette and seven miniature ones.
3e	NTS-SARI-cross	Disney received a full-size Oscar statuette and seven miniature ones**. They** *presented* to him by 10-year-old child actress Shirley Temple.
4a	Original-W, Dress-LS NTS-h1-cross, NTS-SARI-in	At the Voyager 2 images Ophelia appears as an elongated object, the major axis pointing towards Uranus.
4b	SBMT	At the Voyager 2 images Ophelia seems as an elongated object, the main axis *show up on* Uranus.
4c	NTS-h1-in	At the Voyager 2 images Ophelia appears as **a stretched** object, the major axis pointing towards Uranus.
4d	NTS-SARI-cross	At the Voyager 2 images Ophelia, the major axis pointing towards Uranus.
5a	Original-W, NTS-h1-out, SBMT	Graham attended Wheaton College from 1939 to 1943, when he graduated with a BA in anthropology.
5b	NTS-SARI-cross	Graham attended Wheaton College from 1939 to 1943**. He** graduated with a BA in anthropology.
5c	NTS-SARI-in, NTS-h1-in	Graham graduated from Wheaton College from 1939 to 1943.
5d	Dress-LS	Graham **went** to Wheaton College from 1939 to 1943.
6a	Original-W, NTS-h1-any	As a result, although many mosques will not enforce violations, both men and women when attending a mosque must adhere to these guidelines.
6b	Dress-LS	As a result, although many mosques will not enforce violations, both men and women.
6c	NTS-SARI-cross	As a result, many mosques will not enforce violations, both men and women when attending a mosque must **follow** these guidelines.
6d	NTS-SARI-in	As a result, although many mosques will not enforce violations, both men and women when attending a mosque must **stick** to these guidelines.
6e	SBMT	As a result, while many mosques will not *meet the breach*, both men and women when **go to** a mosque must **meet** these guidelines.

Table 8: Simplification examples on the sentences from Wikipedia and Newsela (Newsela, 2016). Correct changes are presented in bold, and the incorrect changes in italics.

based machine translation (SBMT) model. Furthermore, we show that a simple approach to sample multiple hypotheses from a vanilla encoder-decoder can outperform a more complex neural text simplification model tuned with reinforcement learning (Dress-LS), on all human evaluation metrics.

We acknowledge that more work is needed to make sequence-to-sequence models flexible enough for handling out-of-vocabulary words, especially in a cross-domain text simplification. However, neural TS systems were still able to produce grammatical output and correctly model sentence splittings and sentence shortenings even across different text genres.

Our work revealed the challenges that these models face when training and predicting cross-domain, as well as their capacity to correctly perform significant content reduction and improve over the existing text simplification systems.

Acknowledgements

We would like to thank Newsela for making their corpora available, and to Luke Orland for his comments and suggestions. This work has been partially supported by a grant of the Romanian National Authority for Scientific Research and Innovation, CNCS/CCCDI – UEFISCDI, project number PN-III-P2-2.1-53BG/2016, within PNCDI III.

6. Bibliographical References

Amancio, M. A. and Specia, L. (2014). An Analysis of Crowdsourced Text Simplifications . In *Proceedings of the 3rd Workshop on Predicting and Improving Text Readability for Target Reader Populations (PITR)*, pages 123–130.

Bird, S., Klein, E., and Loper, E. (2009). *Natural Language Processing with Python*. O'Reilly Media, Inc., 1st edition.

Bojar, O., Chatterjee, R., Federmann, C., Graham, Y., Haddow, B., Huang, S., Huck, M., Koehn, P., Liu, Q., Logacheva, V., et al. (2017). Findings of the 2017 conference on machine translation (wmt17). In *Proceedings of the Second Conference on Machine Translation*, pages 169–214.

Coster, W. and Kauchak, D. (2011). Learning to Simplify Sentences Using Wikipedia. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 1–9.

Graves, A. (2012). *Supervised sequence labelling with recurrent neural networks*, volume 385. Springer.

Hochreiter, S. and Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, 9(8):1735–1780.

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

Kauchak, D. (2013). Improving text simplification language modeling using unsimplified text data. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1537–1546. ACL.

Klein, G., Kim, Y., Deng, Y., Senellart, J., and Rush, A. M. (2017). OpenNMT: Open-Source Toolkit for Neural Machine Translation. *ArXiv e-prints*.

Luong, M.-T. and Manning, C. D. (2016). Achieving open vocabulary neural machine translation with hybrid word-character models. *arXiv preprint arXiv:1604.00788*.

Luong, T., Pham, H., and Manning, C. D. (2015). Effective approaches to attention-based neural machine translation. In *EMNLP*, pages 1412–1421. The Association for Computational Linguistics.

Newsela. (2016). Newsela article corpus. `https://newsela.com/data`. Version: 2016-01-29.

Nisioi, S., Štajner, S., Ponzetto, S. P., and Dinu, L. P. (2017). Exploring neural text simplification models. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 85–91.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002). BLEU: a method for automatic evaluation of machine translation. In *Proceedings of ACL*.

Sennrich, R., Haddow, B., and Birch, A. (2015). Neural machine translation of rare words with subword units. *arXiv preprint arXiv:1508.07909*.

Specia, L. (2010). Translating from complex to simplified sentences. In *Proceedings of the 9th international conference on Computational Processing of the Portuguese Language (PROPOR)*, volume 6001 of *Lecture Notes in Computer Science*, pages 30–39. Springer Berlin Heidelberg.

Srivastava, N., Hinton, G. E., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R. (2014). Dropout: a simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15(1):1929–1958.

Štajner, S., Bechara, H., and Saggion, H. (2015). A Deeper Exploration of the Standard PB-SMT Approach to Text Simplification and its Evaluation. In *Proceedings of ACL&IJCNLP (Volume 2: Short Papers)*, pages 823–828.

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 the 55th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 97–102.

Štajner, S., Franc-Salvador, M., Rosso, P., and Ponzetto, S. P. (2018). CATS: A Tool for Customised Alignment of Text Simplification Corpora. In *Proceedings of the 11th Language Resources and Evaluation Conference (LREC)*.

Wubben, S., van den Bosch, A., and Krahmer, E. (2012). Sentence simplification by monolingual machine translation. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (ACL): Long Papers - Volume 1*, pages 1015–1024. Association for Computational Linguistics.

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

Zhang, X. and Lapata, M. (2017). Sentence Simplification with Deep Reinforcement Learning. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 584–594.