# One Step Closer to Automatic Evaluation of Text Simplification Systems

Sanja Štajner<sup>1</sup> and Ruslan Mitkov<sup>1</sup> and Horacio Saggion<sup>2</sup>

<sup>1</sup>Research Group in Computational Linguistics, University of Wolverhampton, UK

<sup>2</sup>TALN Research Group, Universitat Pompeu Fabra, Spain

S.Stajner@wlv.ac.uk, R.Mitkov@wlv.ac.uk, horacio.saggion@upf.edu

### Abstract

This study explores the possibility of replacing the costly and time-consuming human evaluation of the grammaticality and meaning preservation of the output of text simplification (TS) systems with some automatic measures. The focus is on six widely used machine translation (MT) evaluation metrics and their correlation with human judgements of grammaticality and meaning preservation in text snippets. As the results show a significant correlation between them, we go further and try to classify simplified sentences into: (1) those which are acceptable; (2) those which need minimal post-editing; and (3) those which should be discarded. The preliminary results, reported in this paper, are promising.

### 1 Introduction

Lexically and syntactically complex sentences can be difficult to understand for non-native speakers (Petersen and Ostendorf, 2007; Aluísio et al., 2008b), and for people with language impairments, e.g. people diagnosed with aphasia (Carroll et al., 1999; Devlin, 1999), autism spectrum disorder (Štajner et al., 2012; Martos et al., 2012), dyslexia (Rello, 2012), congenital deafness (Inui et al., 2003), and intellectual disability (Feng, 2009). At the same time, long and complex sentences are also a stumbling block for many NLP tasks and applications such as parsing, machine translation, information retrieval, and summarisation (Chandrasekar et al., 1996). This justifies the need for Text Simplification (TS) systems which would convert such sentences into their simpler and easier-to-read variants, while at the same time preserving the original meaning.

So far, TS systems have been developed for English (Siddharthan, 2006; Zhu et al., 2010; Wood-

send and Lapata, 2011a; Coster and Kauchak, 2011; Wubben et al., 2012), Spanish (Saggion et al., 2011), and Portuguese (Aluísio et al., 2008a), with recent attempts at Basque (Aranzabe et al., 2012), Swedish (Rybing et al., 2010), Dutch (Ruiter et al., 2010), and Italian (Barlacchi and Tonelli, 2013).

Usually, TS systems are either evaluated for: (1) the quality of the generated output, or (2) the effectiveness/usefulness of such simplification on reading speed and comprehension of the target population. For the purpose of this study we focused only on the former. The quality of the output generated by TS systems is commonly evaluated by using a combination of readability metrics (measuring the degree of simplification) and human assessment (measuring the grammaticality and meaning preservation). Despite the noticeable similarity between evaluation of the fluency and adequacy of a machine translation (MT) output, and evaluation of grammaticality and meaning preservation of a TS system output, there have been no works exploring whether any of the MT evaluation metrics are well correlated with the latter, and could thus replace the time-consuming human assessment.

The contributions of the present work are the following:

- It is the first study to explore the possibility of replacing human assessment of the quality of TS system output with automatic evaluation.
- It is the first study to investigate the correlation of human assessment of TS system output with MT evaluation metrics.
- It proposes a decision-making procedure for the classification of simplified sentences into: (1) those which are acceptable; (2) those which need further post-editing; and (3) those which should be discarded.

#### 2 Related Work

The output of the TS system proposed by Siddharthan (2006) was rated for grammaticality and meaning preservation by three human evaluators. Similarly, Drndarevic et al. (2013) evaluated the grammaticality and the meaning preservation of automatically simplified Spanish sentences on a Likert scale with the help of twenty-five human annotators. Additionally, the authors used seven readability metrics to assess the degree of simplification. Woodsend and Lapata (2011b), and Glavaš and Štajner (2013) used human annotators' ratings for evaluating simplification, meaning preservation, and grammaticality, while additionally applying several readability metrics for evaluating complexity reduction in entire texts.

Another set of studies approached TS as an MT task translating from "original" to "simplified" language, e.g. (Specia, 2010; Woodsend and Lapata, 2011a; Zhu et al., 2010). In this case, the quality of the output generated by the system was evaluated using several standard MT evaluation metrics: BLEU (Papineni et al., 2002), NIST (Doddington, 2002), and TERp (Snover et al., 2009).

### 3 Methodology

All experiments were conducted on a freely available sentence-level dataset<sup>1</sup>, fully described in (Glavaš and Štajner, 2013), and the two datasets we derived from it. The original dataset and the instructions for the human assessment are given in the next two subsections. Section 3.3 explains how we derived two additional datasets from the original one, and to what end. Section 3.4 describes the automatic MT evaluation metrics used as features in correlation and classification experiments; Section 3.5 presents the main goals of the study; and Section 3.6 describes the conducted experiments.

#### 3.1 Original dataset

The dataset contains 280 pairs of original sentences and their corresponding simplified versions annotated by humans for grammaticality, meaning preservation, and simplicity of the simplified version. We used all sentence pairs, focusing only on four out of eight available features: (1) the original text, (2) the simplified text, (3) the grammaticality score, and (4) the score for meaning preservation.<sup>2</sup>

Category	weighted $\kappa$	Pearson	MAE
Grammaticality	0.68	0.77	0.18
Meaning	0.53	0.67	0.37
Simplicity	0.54	0.60	0.28

Table 1: IAA from (Glavaš and Štajner, 2013)

The simplified versions of original sentences were obtained by using four different simplification methods: baseline, sentence-wise, eventwise, and pronominal anaphora. The baseline retains only the main clause of a sentence, and discards all subordinate clauses, based on the output of the Stanford constituency parser (Klein and Manning, 2003). Sentence-wise simplification eliminates all those tokens in the original sentence that do not belong to any of the extracted factual event mentions, while the event-wise simplification transforms each factual event mention into a separate sentence of the output. The last simplification scheme (pronominal anaphora) additionally employs pronominal anaphora resolution on top of the event-wise simplification scheme.<sup>3</sup>

#### 3.2 Human Assessment

Human assessors were asked to score the given sentence pairs (or text snippets in the case of split sentences) on a 1-3 scale based on three criteria: Grammaticality (1 - ungrammatical, 2 - minor problems with grammaticality, 3 - grammatical), Meaning (1 – meaning is seriously changed or most of the relevant information lost, 2 - some of the relevant information is lost but the meaning of the remaining information is unchanged, 3 – all relevant information is kept without any change in meaning), and Simplicity (1 - a lot of irrelevant information is retained, 2 - some of irrelevant information is retained, 3 – all irrelevant information is eliminated). The inter-annotator agreement (IAA) was calculated using weighted Kappa (weighted  $\kappa$ ), Pearson's correlation (*Pearson*), and mean average error (MAE), and the obtained results are presented in Table 1. A few examples of assigned scores are given in Table 2, where G, M, and Sdenote human scores for grammaticality, meaning preservation and simplicity respectively.

<sup>&</sup>lt;sup>1</sup>http://takelab.fer.hr/data/evsimplify/

<sup>&</sup>lt;sup>2</sup>The other four features contain the pairID, groupID, the method with which the simplification was obtained, and the

score for simplicity, which are not relevant here.

<sup>&</sup>lt;sup>3</sup>For more detailed explanation of simplification schemes and the dataset see (Glavaš and Štajner, 2013).

Ex.	Original	Simplified	G	М	S	SM
(a)	"It is understood the dead girl had	"The dead girl had been living at	3	3	3	S
	been living at her family home, in	her family home, in a neighbouring				
	a neighbouring housing estate, and	housing estate and was visiting her				
	was visiting her older sister at the	older sister."				
	time of the shooting."					
(b)	"On Facebook, more than 10,000	"On Facebook, more than 10,000	2	3	3	S
	people signed up to a page an-	people signed to a page announcing				
	nouncing an opposition rally for	an opposition rally for Saturday."				
	Saturday."		-		-	a
(c)	"Joel Elliott, also 22, of North	"Joel Elliott was charged on May 3	3	2	3	S
	Road, Brighton, was charged on	with murder. He appeared at Lewes				
	May 3 with muraer. He appeared at	Crown Court on May 8.				
	Lewes Crown Court on May 8 but did not enter a plag."					
(d)	"For years the former Bosnia Serb	"For years the former Bosnia Serb	1	2	3	S
(u)	army commander Ratko Mladic had	army commander Ratko Mladic had	1	2	5	5
	evaded capture and was one of the	evaded but his time the run ended				
	world's most wanted men, but his	last year he was arrested near Bel-				
	time on the run finally ended last	grade."				
	year when he was arrested near	0				
	Belgrade."					
(e)	"Police have examined the scene at	"Police have examined the scene	3	1	3	Р
	a house at William Court in Bel-	at William Court near Magherafelt.				
	laghy, near Magherafelt for clues to	The incident has stunned the com-				
	the incident which has stunned the	munity."				
	community."					
(f)	"Rastan, 25 km (15 miles) north	"Rastan has slipped government	2	1	3	Р
	of Homs city, has slipped in and	control several times. The uprising				
	out of government control several	erupted in March 2011."				
	times since the uprising against As-					
	sad erupted in March 2011."	((D))	1	1	2	D
(g)	"But opposition parties and inter-	"But opposition parties and inter-	1	I	3	В
	national observers said the vote was	national observers sala .				
	alleged ballot box stuffing and false					
	voter rolls"					
(h)	"Foreign Affairs Secretary Albert	"Foreign Affairs Secretary Albert	3	3	1	В
	del Rosario was seeking a diplo-	del Rosario was seeking a diplo-		-	-	-
	matic solution with Chinese Ambas-	matic solution with Chinese Ambas-				
	sador Ma Keqing, the TV network	sador Ma Keqing, the TV network				
	said."	said."				
(h)	" On Wednesday, two video jour-	"On Wednesday two video journal-	3	2	2	Е
	nalists working for the state-owned	ists were briefly detained outside				
	RIA Novosti news agency were	the Election Commission building.				
	briefly detained outside the Election	Two video journalists worked for				
	Commission building where Putin	the state-owned RIA Novosti news				
	was handing in his application to	agency. Putin was handing in his				
	run."	application."				

Table 2: Human evaluation examples (*G*, *M*, and *S* correspond to the human scores for grammaticality, meaning preservation and simplicity, and *SM* denotes the simplification method used: B – baseline, S – sentence-wise, E – event-wise, and P – pronominal anaphora)

#### 3.3 Derived Datasets

The original dataset (*Original*) contains separate scores for grammaticality (G), meaning preservation (M), and simplicity (S), each of them on a 1–3 scale. From this dataset we derived two additional ones: *Total3* and *Total2*.

The Total3 dataset contains three marks (OK – use as it is, PE - post-editing required, and Dis - discard) derived from G and M in the Original dataset. Those simplified sentences which scored '3' for both meaning preservation (M) and grammaticality (G) are placed in the OK class as they do not need any kind of post-editing. A closer look at the remaining sentences suggests that any simplified sentence which got a score '2' or '3' for meaning preservation (M) could be easily postedited, i.e. it requires minimal changes which are obvious from its comparison to the corresponding original. For instance, in the sentence (b) in Table 2 the only change that needs to be made is adding the word "up" after "signed". Those sentences which scored '2' for meaning need slightly more, albeit simple modification. The simplified text snippet (c) in Table 2 would need "but did not enter a plea" added at the end of the last sentence. The next sentence (d) in the same table needs a few more changes, but still very minor ones: adding the word "capture" after "had evaded", adding the preposition "on" before "the run", and adding "when" after "last year". Therefore, we grouped all those sentences into one class - PE (sentences which require a minimal postediting effort). Those sentences which scored '1' for meaning need to either be left in their original form or simplified from scratch. We thus classify them as Dis. This newly created dataset (Total3) allows us to investigate whether we could automatically classify simplified sentences into those three categories, taking into account both grammaticality and meaning preservation at the same time.

The *Total2* dataset contains only two marks ('0' and '1') which correspond to the sentences which should be discarded ('0') and those which should be retained ('1'), where '0' corresponds to *Dis* in *Total3*, and '1' corresponds to the union of *OK* and *PE* in *Total3*. The derivation procedure for both datasets is presented in Table 3. We wanted to investigate whether the classification task would be simpler (better performed) if there were only two classes instead of three. In the case that such classification task would be simpler (better performed) if there were only two classes instead of three. In the case that such classification task would be simpler (better performed) if there were only two classes instead of three.

sification could be performed with satisfactory accuracy, all sentences classified as '0' would be left in their original form or simplified with some different simplification strategy, while those classified as '1' would be sent for a quick human postediting procedure.

Ori	ginal	Total2	Total2
G	Μ	10(1)	101812
3	3	OK	1
2	3	PE	1
1	3	PE	1
3	2	PE	1
2	2	PE	1
1	2	PE	1
3	1	Dis	0
2	1	Dis	0
1	1	Dis	0

#### Table 3: Datasets

Here it is important to mention that we decided not to use human scores for simplicity (S) for several reasons. First, simplicity was defined as the amount of irrelevant information which was eliminated. Therefore, we cannot expect that any of the six MT evaluation metrics would have a significant correlation with this score (except maybe TERp and, in particular, one of its parts - 'number of deletions'. However, none of the two demonstrated any significant correlation with the simplicity score, and those results are thus not reported in this paper). Second, the output sentences with a low simplicity score are not as detrimental for the TS system as those with a low grammaticality or meaning preservation score. The sentences with a low simplicity score would simply not help the target user read faster or understand better, but would not do any harm either. Alternatively, if the target "user" is an MT or information extraction (IE) system, or a parser for example, such sentences would not lower the performance of the system; they would just not improve it. Low scores for G and M, however, would lead to a worse performance for such NLP systems, longer reading time, and a worse or erroneous understanding of the text. Third, the simplicity of the output (or complexity reduction performed by a TS system) could be evaluated separately, in a fully automatic manner - using some readability measures or average sentence length as features (as in (Drndarević et al., 2013; Glavaš and Štajner, 2013) for example).

#### 3.4 Features: MT Evaluation Metrics

In all experiments, we focused on six commonly used MT evaluation metrics. These are cosine similarity (using the bag-of-words representation), METEOR (Denkowski and Lavie, 2011), TERp (Snover et al., 2009), TINE (Rios et al., 2011), and two components of TINE: T-BLEU (which differs from the standard BLEU (Papineni et al., 2002) by using 3-grams, 2-grams, and 1-grams when there are no 4-grams found, where the "original" BLEU would give score '0') and SRL (which is the component of TINE based on semantic role labeling using SENNA<sup>4</sup>). Although these two components contribute equally to TINE (thus being linearly correlated with TINE), we wanted to investigate which one of them contributes more to the correlation of TINE with human judgements. Given their different natures, we expect T-BLEU to contribute more to the correlation of TINE with human judgements of grammaticality, and SRL to contribute more to the correlation of TINE with human judgements of meaning preservation.

As we do not have the reference for the simplified sentence, all metrics are applied in a slightly different way than in MT. Instead of evaluating the translation hypothesis (output of the automatic TS system in our case) with the corresponding reference translation (which would be a 'gold standard' simplified sentence), we apply the metrics to the output of the automatic TS system comparing it with the corresponding original sentence. Given that the simplified sentences in the used dataset are usually shorter than the original ones (due to the elimination of irrelevant content which was the main focus of the TS system proposed by Glavaš and Štajner (2013)), we expect low scores of T-BLEU and METEOR which apply a brevity penalty. However, our dataset does not contain any kind of lexical simplification, but rather copies all relevant information from the original sentence<sup>5</sup>. Therefore, we expect the exact matches of word forms and semantic role labels (which are components of the MT evaluation metrics) to have a good correlation to human judgements of grammaticality and meaning preservation.

### 3.5 Goal

After we obtained the six automatic metrics (cosine, METEOR, TERp, TINE, T-BLEU, and SRL), we performed two sets of experiments, trying to answer two main questions:

- 1. Are the chosen MT evaluation metrics correlated with the human judgements of grammaticality and meaning preservation of the TS system output?
- 2. Could we automatically classify the simplified sentences into those which are: (1) correct, (2) require a minimal post-editing, (3) incorrect and need to be discarded?

A positive answer to the first question would mean that there is a possibility of finding an automatic metric (or a combination of several automatic metrics) which could successfully replace the time consuming human evaluation. The search for that "ideal" combination of automatic metrics could be performed by using various classification algorithms and carefully designed features. If we manage to classify simplified sentences into the three aforementioned categories with a satisfying accuracy, the benefits would be two-fold. Firstly, such a classification system could be used for an automatic evaluation of TS systems and an easy comparison of their performances. Secondly, it could be used inside a TS system to mark those sentences of low quality which need to be checked further, or those sentences whose original meaning changed significantly. The latter could then be left in their original form or simplified using some different technique.

### 3.6 Experiments

The six experiments conducted in this study are presented in Table 4. The first two experiments had the aim of answering the first question (Section 3.5) as to whether the chosen MT metrics correlate with the human judgements of grammaticality (G) and meaning preservation (M) of the TS system output. The results were obtained in terms of Pearson's, Kendall's and Spearman's correlation coefficients. The third and the fourth experiments (Table 4) could be seen as the intermediate experiments exploring the possibility of automatic classification of simplified sentences according to their grammaticality, and meaning preservation. The main experiment was the fifth experiment, trying to answer the second question (Section 3.5)

<sup>&</sup>lt;sup>4</sup>http://ml.nec-labs.com/senna/

<sup>&</sup>lt;sup>5</sup>The exceptions being changes of gerundive forms into past tense, and anaphoric pronoun resolution in some simplification schemes. See Section 3.1 and (Glavaš and Štajner, 2013) for more details.

Exp.	Description
1.	Correlation of the six automatic MT metrics with the human scores for <i>Grammaticality</i>
2.	Correlation of the six automatic MT metrics with the human scores for <i>Meaning</i> preservation
3.	Classification of the simplified sentences into 3 classes ('1' – Bad, '2' – Medium, and '3' – Good) according to
	their Grammaticality
4.	Classification of the simplified sentences into 3 classes ('1' – Bad, '2' – Medium, and '3' – Good) according to
	their <i>Meaning</i> preservation
5	Classification of the simplified sentences into 3 classes ( $OK$ PF Dis) according to their Total3 score

6. Classification of the simplified sentences into 2 classes ('1' – *Retain*, '0' – *Discard*) according to their *Total2* score

Table 4: Experiments

as to whether we could automatically classify the simplified sentences into those which are: (1) correct (OK), (2) require minimal post-editing (PE), and (3) incorrect and need to be discarded (Dis). The last experiment (Table 4) was conducted with the aim of exploring whether the classification of simplified sentences into only two classes – *Retain* (for further post-editing) and *Discard* – would lead to better results than the classification into three classes (OK, PE, and Dis) in the fifth experiment.

All classification experiments were performed in Weka workbench (Witten and Frank, 2005; Hall et al., 2009), using seven classification algorithms in a 10-fold cross-validation setup:

- NB NaiveBayes (John and Langley, 1995),
- SMO Weka implementation of Support Vector Machines (Keerthi et al., 2001) with normalisation (n) or with standardisation (s),
- Logistic (le Cessie and van Houwelingen, 1992),
- Lazy.IBk K-nearest neighbours (Aha and Kibler, 1991),
- JRip a propositional rule learner (Cohen, 1995),
- J48 Weka implementation of C4.5 (Quinlan, 1993).

As a baseline we use the classifier which assigns the most frequent (majority) class to all instances.

## 4 Results and Discussion

The results of the first two experiments (correlation experiments in Table 4) are presented in Section 4.1, while the results of the other four experiments (classification experiments in Table 4) can be found in Section 4.2. When interpreting the results of all experiments, it is important to keep in mind that human agreements for meaning preservation (M) and grammaticality (G) were acceptable but far from perfect (Section 3.2), and thus it would be unrealistic to expect the correlation between the MT evaluation metrics and human judgements or the agreement of the classification system with human assessments to be higher than the reported IAA agreement.

### 4.1 Correlation of Automatic Metrics with Human Judgements

The correlations of automatic metrics with human judgements of grammaticality and meaning preservation are given in Tables 5 and 6 respectively. Statistically significant correlations (at a 0.01 level of significance) are presented in bold.

Metric	Pearson	Kendall	Spearman
cosine	0.097	0.092	0.115
METEOR	0.176	0.141	0.178
T-BLEU	0.226	0.185	0.234
SRL	0.097	0.076	0.095
TINE	0.175	0.145	0.181
TERp	-0.208	-0.158	-0.198

 Table 5: Correlation between automatic evaluation

 metrics and human scores for grammaticality

Metric	Pearson	Kendall	Spearman
cosine	0.293	0.262	0.334
METEOR	0.386	0.322	0.405
T-BLEU	0.442	0.382	0.475
SRL	0.348	0.285	0.356
TINE	0.427	0.385	0.447
TERp	-0.414	-0.336	-0.416

Table 6: Correlation between automatic evaluation metrics and human scores for meaning preservation

It can be noted that human perception of grammaticality is positively correlated with three auto-

Algorithm	Gra	mmatic	ality	1	Meaning		Total3			Total2		
Algorium	P	R	F	P	R	F	Р	R	F	P	R	F
NB	0.53	0.46	0.48	0.54	0.54	0.54	0.54	0.53	0.53	0.74	0.69	0.71
SMO(n)	0.39	0.63	0.48	0.52	0.49	0.45	0.43	0.53	0.44	0.55	0.74	0.63
SMO(s)	0.39	0.63	0.48	0.57	0.56	0.55	0.57	0.55	0.51	0.60	0.73	0.63
Logistic	0.45	0.61	0.49	0.57	0.57	0.56	0.61	0.60	0.59	0.75	0.77	0.74
Lazy.IBk	0.57	0.58	0.57	0.50	0.50	0.50	0.54	0.54	0.54	0.73	0.73	0.73
JRip	0.41	0.59	0.48	0.53	0.50	0.48	0.57	0.56	0.55	0.72	0.75	0.73
J48	0.45	0.61	0.49	0.48	0.47	0.47	0.59	0.57	0.54	0.68	0.71	0.69
baseline	0.39	0.63	0.48	0.17	0.41	0.24	0.21	0.46	0.29	0.55	0.74	0.63

Table 7: Classification results (the best performances are shown in bold; baseline uses the majority class)

Actual	Grar	nmatica	lity	Meaning			
Actual	Good	Med.	Bad	Good	Med.	Bad	
Good	127	21	23	50	31	7	
Med.	29	19	10	24	73	16	
Bad	24	9	10	9	31	31	

Table 8: Confusion matrices for the best classifications according to *Grammaticality* (Lazy.IBk) and *Meaning* (Logistic). The number of "severe" classification mistakes (classifying *Good* as *Bad* or vice versa) are presented in bold.

matic measures - METEOR, T-BLEU, and TINE, while it is negatively correlated with TERp (TERp measures the number of edits necessary to perform on the simplified sentence to transform it into its original one, i.e. the higher the value of TERp, the less similar the original and its corresponding simplified sentence are. The other five MT metrics measure the similarity between the original and its corresponding simplified version, i.e. the higher their value is, the more similar are the sentences are). All the MT metrics appear to be even better correlated with the human scores for meaning preservation (Table 6), demonstrating six positive and one (TERp) negative statistically significant correlation with M. The correlation is the highest for T-BLEU, TINE, and TERp, though closely followed by all others.

### 4.2 Sentence Classification

The results of the four classification experiments (Section 3.6) are given in Table 7.

At first glance, the performance of the classification algorithms seems similar for the first two tasks (classification of the simplified sentences according to their *Grammaticality* and *Meaning* preservation). However, one needs to take into account that the baseline for the first task was much much higher than for the second task (Table 7). Furthermore, it can be noted that for the first task, recall was significantly higher than precision for most classification algorithms (all except NB and Logistic), while for the second task they were very similar in all cases. More importantly, a closer look at the confusion matrices reveals that most of the incorrectly classified sentences were assigned to the nearest class (*Medium* into *Bad* or *Good*; *Bad* into *Medium*; and *Good* into *Medium*<sup>6</sup>) in the second task, while it was not the case in the first task (Table 8).

Classification performed on the *Total3* dataset outperformed both previous classifications – that based on *Grammaticality* and that based on *Meaning* – on four different algorithms (NB, Logistic, JRip, and J48). Classification conducted on *Total3* using Logistic outperformed all results of classifications on either *Grammaticality* or *Meaning* separately (Table 7). It reached a 0.61, 0.60, and 0.59 score for the weighted precision (P), recall (R), and F-measure (F), respectively, thus outperforming the baseline significantly. More importantly, classification on the *Total3* dataset led to significantly fewer mis-classifications between *Good* and *Bad* (Table 9) than the classification based on *Grammaticality*, and slightly less than

<sup>&</sup>lt;sup>6</sup>*Bad, Medium,* and *Good* correspond to marks '1', '2', and '3' given by human evaluators.

Actual		Total3				
Actual	OK	PE	Dis.			
OK	41	32	4			
PE	17	85	12			
Dis.	6	31	28			

Table 9: Confusion matrix for the best classification according to *Total3* (Logistic). The number of "severe" classification mistakes (classifying *Good* as *Bad* or vice versa) are presented in bold.

Actual	Total2				
Actual	Retain	Discard			
Retain	21	50			
Discard	12	189			

Table 10: Confusion matrix for the best classification according to *Total2* (Logistic). The number of "severe" classification mistakes (classifying *Retain* as *Discard* or vice versa) are presented in bold.

the classification based only on *Meaning* (Table 8). Therefore, it seems that simplified sentences are better classified into three classes giving a unique score for both grammaticality and preservation of meaning together.

The binary classification experiments based on the *Total2* led to results which significantly outperformed the baseline in terms of precision and F-measure (Table 7). However, they resulted in a great number of sentences which should be retained (*Retain*) being classified into those which should be discarded (*Discard*) and vice versa (Table 10). Therefore, it seems that it would be better to opt for classification into three classes (*Total3*) than for classification into two classes (*Total2*).

Additionally, we used CfsSubsetEval attribute selection algorithm (Hall and Smith, 1998) in order to identify the 'best' subset of features. The 'best' subsets of features for each of the four classification tasks returned by the algorithm are listed in Table 11. However, the classification performances achieved (P, R, and F) when using only the 'best' features did not differ significantly from those when using all initially selected features, and thus are not presented in this paper.

### 5 Limitations

The used dataset does not contain any kind of lexical simplification (Glavaš and Štajner, 2013).

Classification	'Best' features
Meaning	{TERp, T-BLEU, SRL, TINE}
Grammaticality	{TERp, T-BLEU}
New3	{TERp, T-BLEU, SRL, TINE}
New2	$\{\text{TERp}, \text{T-BLEU}, \text{SRL}\}$

Table 11: The 'best' features (CfsSubsetEval)

Therefore, one should consider the limitation of this TS system which performs only syntactic simplification and content reduction. On the other hand, the dataset used contains a significant content reduction in most of the sentences. If the same experiments were conducted on a dataset which performs only syntactic simplification, we would expect much higher correlation of MT evaluation metrics to human judgements, due to the lesser impact of the brevity penalty in that case.

If we were to apply the same MT evaluation metrics to a TS system which additionally performs some kind of lexical simplification (either a simple lexical substitution or paraphrasing), the correlation results for T-BLEU and cosine similarity would be lower (due to the lower number of exact matches), but not for METEOR, TERp and SRL (and thus TINE as well). As a similar problem is also present in the evaluation of MT systems where the obtained output could differ from the reference translation (while still being equally good), METEOR, TERp, and SRL in TINE additionally use inexact matching. The first two use the stem, synonym, and paraphrase matches, while SRL uses ontologies and thesaurus.

### 6 Conclusions and Future Work

While the results reported are preliminary and their universality needs to be validated on different TS datasets, the experiments and results presented can be regarded as a promising step towards an automatic assessment of grammaticality and meaning preservation for the output of TS systems. In addition and to the best of our knowledge, there are no such datasets publicly available other than the one used. Nevertheless, we hope that these results would initiate an interesting discussion in the TS community and start a new direction of studies towards automatic evaluation of text simplification systems.

### Acknowledgements

The research described in this paper was partially funded by the European Commission under the Seventh (FP7-2007-2013) Framework Programme for Research and Technological Development (FP7-ICT-2011.5.5 FIRST 287607).

### References

- D. Aha and D. Kibler. 1991. Instance-based learning algorithms. *Machine Learning*, 6:37–66.
- S. M. Aluísio, L. Specia, T. A. S. Pardo, E. G. Maziero, H. M. Caseli, and R. P. M. Fortes. 2008a. A corpus analysis of simple account texts and the proposal of simplification strategies: first steps towards text simplification systems. In *Proceedings of the* 26th annual ACM international conference on Design of communication, SIGDOC '08, pages 15–22, New York, NY, USA. ACM.
- S. M. Aluísio, L. Specia, T. A.S. Pardo, E. G. Maziero, and R. P.M. Fortes. 2008b. Towards brazilian portuguese automatic text simplification systems. In *Proceedings of the eighth ACM symposium on Document engineering*, DocEng '08, pages 240–248, New York, NY, USA. ACM.
- M. J. Aranzabe, A. Díaz De Ilarraza, and I. González. 2012. First Approach to Automatic Text Simplification in Basque. In *Proceedings of the first Natural Language Processing for Improving Textual Accessibility Workshop (NLP4ITA).*
- G. Barlacchi and S. Tonelli. 2013. ERNESTA: A sentence simplification tool for childrens stories in italian. In *Computational Linguistics and Intelligent Text Processing*.
- J. Carroll, G. Minnen, D. Pearce, Y. Canning, S. Devlin, and J. Tait. 1999. Simplifying text for language-impaired readers. In *Proceedings of the* 9th Conference of the European Chapter of the ACL (EACL'99), pages 269–270.
- R. Chandrasekar, Christine Doran, and B. Srinivas. 1996. Motivations and methods for text simplification. In *In Proceedings of the Sixteenth International Conference on Computational Linguistics* (COLING '96, pages 1041–1044.
- W. Cohen. 1995. Fast Effective Rule Induction. In Proceedings of the Twelfth International Conference on Machine Learning, pages 115–123.
- W. Coster and D. Kauchak. 2011. Learning to Simplify Sentences Using Wikipedia. In *Proceedings* of the 49th Annual Meeting of the Association for Computational Linguistics, pages 1–9.

- M. Denkowski and A. Lavie. 2011. Meteor 1.3: Automatic Metric for Reliable Optimization and Evaluation of Machine Translation Systems. In *Proceedings of the EMNLP Workshop on Statistical Machine Translation*.
- S. Devlin. 1999. *Simplifying natural language text for aphasic readers*. Ph.D. thesis, University of Sunderland, UK.
- G. Doddington. 2002. Automatic evaluation of machine translation quality using n-gram coocurrence statistics. In *Proceedings of the second international conference on Human Language Technology Research*, pages 138–145. Morgan Kaufmann Publishers Inc.
- B. Drndarević, S. Štajner, S. Bott, S. Bautista, and H. Saggion. 2013. Automatic Text Simplication in Spanish: A Comparative Evaluation of Complementing Components. In Proceedings of the 12th International Conference on Intelligent Text Processing and Computational Linguistics. Lecture Notes in Computer Science. Samos, Greece, 24-30 March, 2013., pages 488–500.
- L. Feng. 2009. Automatic readability assessment for people with intellectual disabilities. In *SIGACCESS Access. Comput.*, number 93, pages 84–91. ACM, New York, NY, USA, jan.
- G. Glavaš and S. Štajner. 2013. Event-Centered Simplication of News Stories. In Proceedings of the Student Workshop held in conjunction with RANLP 2013, Hissar, Bulgaria, pages 71–78.
- M. A. Hall and L. A. Smith. 1998. Practical feature subset selection for machine learning. In C. Mc-Donald, editor, *Computer Science '98 Proceedings* of the 21st Australasian Computer Science Conference ACSC'98, pages 181–191. Berlin: Springer.
- M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten. 2009. The weka data mining software: an update. *SIGKDD Explor. Newsl.*, 11:10–18, November.
- K. Inui, A. Fujita, T. Takahashi, R. Iida, and T. Iwakura. 2003. Text simplification for reading assistance: a project note. In *Proceedings of the second international workshop on Paraphrasing - Volume 16*, PARAPHRASE '03, pages 9–16, Stroudsburg, PA, USA. Association for Computational Linguistics.
- G. H. John and P. Langley. 1995. Estimating Continuous Distributions in Bayesian Classifiers. In *Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence*, pages 338–345.
- S. S. Keerthi, S. K. Shevade, C. Bhattacharyya, and K. R. K. Murthy. 2001. Improvements to Platt's SMO Algorithm for SVM Classifier Design. *Neural Computation*, 13(3):637–649.

- D. Klein and C.D. Manning. 2003. Accurate unlexicalized parsing. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics*, volume 1, pages 423–430. Association for Computational Linguistics.
- S. le Cessie and J.C. van Houwelingen. 1992. Ridge Estimators in Logistic Regression. *Applied Statistics*, 41(1):191–201.
- J. Martos, S. Freire, A. González, D. Gil, and M. Sebastian. 2012. D2.1: Functional requirements specifications and user preference survey. Technical report, FIRST technical report.
- K. Papineni, S. Roukos, T. Ward, and W. Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of ACL*.
- S. E. Petersen and M. Ostendorf. 2007. Text Simplification for Language Learners: A Corpus Analysis. In *Proceedings of Workshop on Speech and Language Technology for Education*.
- R. Quinlan. 1993. C4.5: Programs for Machine Learning. Morgan Kaufmann Publishers, San Mateo, CA.
- L. Rello. 2012. Dyswebxia: a model to improve accessibility of the textual web for dyslexic users. In SIGACCESS Access. Comput., number 102, pages 41–44. ACM, New York, NY, USA, January.
- M. Rios, W. Aziz, and L. Specia. 2011. TINE: A metric to assess MT adequacy. In *Proceedings of the Sixth Workshop on Statistical Machine Translation* (WMT-2011), Edinburgh, UK.
- M. B. Ruiter, T. C. M. Rietveld, Cucchiarini C., Krahmer E. J., and H. Strik. 2010. Human Language Technology and communicative disabilities: Requirements and possibilities for the future. In *Proceedings of the the seventh international conference on Language Resources and Evaluation (LREC).*
- J. Rybing, C. Smithr, and A. Silvervarg. 2010. Towards a Rule Based System for Automatic Simplification of Texts. In *The Third Swedish Language Technology Conference*.
- H. Saggion, E. Gómez Martínez, E. Etayo, A. Anula, and L. Bourg. 2011. Text Simplification in Simplext: Making Text More Accessible. *Revista de la Sociedad Española para el Procesamiento del Lenguaje Natural*, 47:341–342.
- A. Siddharthan. 2006. Syntactic simplification and text cohesion. *Research on Language & Computation*, 4(1):77–109.
- M. Snover, N. Madnani, B. Dorr, and R. Schwartz. 2009. Fluency, Adequacy, or HTER? Exploring Different Human Judgments with a Tunable MT Metric. In Proceedings of the Fourth Workshop on Statistical Machine Translation at the 12th Meeting of the European Chapter of the Association for Computational Linguistics (EACL-2009), Athens, Greece.

- L. Specia. 2010. Translating from complex to simplified sentences. In *Proceedings of the 9th international conference on Computational Processing of the Portuguese Language*, pages 30–39, Berlin, Heidelberg.
- S. Štajner, R. Evans, C. Orasan, and R. Mitkov. 2012. What Can Readability Measures Really Tell Us About Text Complexity? In Proceedings of the LREC'12 Workshop: Natural Language Processing for Improving Textual Accessibility (NLP4ITA), Istanbul, Turkey.
- I. H. Witten and E. Frank. 2005. *Data mining: practical machine learning tools and techniques*. Morgan Kaufmann Publishers.
- K. Woodsend and M. Lapata. 2011a. Learning to Simplify Sentences with Quasi-Synchronous Grammar and Integer Programming. In Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing (EMNLP).
- K. Woodsend and M. Lapata. 2011b. WikiSimple: Automatic Simplification of Wikipedia Articles. In *Proceedings of the 25th AAI Coference on Artificial Intelligence*.
- S. Wubben, A. van den Bosch, and E. Krahmer. 2012. Sentence simplification by monolingual machine translation. In Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers - Volume 1, ACL '12, pages 1015– 1024, Stroudsburg, PA, USA. Association for Computational Linguistics.
- Z. Zhu, D. Berndard, and I. Gurevych. 2010. A Monolingual Tree-based Translation Model for Sentence Simplification. In *Proceedings of the 23rd International Conference on Computational Linguistics* (Coling 2010), pages 1353–1361.