Automatic Assessment of Absolute Sentence Complexity

Sanja Štajner, Simone Paolo Ponzetto and **Heiner Stuckenschmidt**
Data and Web Science Group, University of Mannheim, Germany
{sanja,simone,heiner}@informatik.uni-mannheim.de

Abstract

Lexically and syntactically simpler sentences result in shorter reading time and better understanding in many people. However, no reliable systems for automatic assessment of sentence complexity have been proposed so far. Instead, the assessment is usually done manually, requiring expert human annotators. To address this problem, we first define the sentence complexity assessment as a five-level classification task, and build a 'gold standard' dataset. Next, we propose robust systems for sentence complexity assessment, using a novel set of features based on leveraging lexical properties of freely available corpora, and investigate the impact of the feature type and corpus size on the classification performance.

1 Introduction

Lexically or syntactically complex sentences can be difficult to understand for children [De Belder and Moens, 2010], non-native speakers [Petersen and Ostendorf, 2007], people with low literacy [Aluísio and Gasperin, 2010] or various kinds of reading or cognitive impairments [Carroll *et al.*, 1999; Saggion *et al.*, 2015]. Sentence simplification aims at transforming such sentences into variants which are simpler to understand for the target population. In this process, regardless of whether it is done manually or automatically, it is important to: (T1) first identify sentences which might be complex and simplify only them (especially in the case of automatic text simplification, as automatic simplification of already simple sentences can lead to a more complex output[1]); (T2) identify the simplest or simple enough (i.e. the easiest –or easy enough– to understand) variant for the target population among several possible (manual or automatic) simplifications. Both tasks are still done manually by expert human annotators, which makes it expensive and time-consuming, and significantly limits the field of text simplification and text accessibility.

To the best of our knowledge, the first task (T1) has not been addressed so far (although the related task of complex word identification (CWI) has been addressed in the SemEval-2016 CWI shared task[2]). The second task (T2) has been tackled in the QATS shared task [Štajner *et al.*, 2016b], where the participants were provided with 505 sentence pairs for training and 126 sentence pairs for testing, both marked (among other marks for grammaticality, meaning preservation and overall quality) for their simplicity as *bad* (difficult to understand), *OK* (somewhat difficult to understand), or *good* (easy to understand). All sentence pairs were pairs of an original sentence and their automatic simplifications by several automatic text simplification (ATS) systems. This limited the application of the proposed systems to assessing only the output of automatic text simplification systems, and not original sentences or manually simplified ones. Furthermore, a three-level scale might not be enough to account for differences among many variants of the same sentence.

To address all those issues, we first compile a 'gold standard' evaluation dataset (Section 3) which contains a high variety of sentences (original sentences, manual simplifications, and automatic simplifications) human evaluated for their simplicity on a 1–5 level scale (enabling thus a more fine-grained complexity assessment). By having all these three types of sentences, our dataset –unlike the one offered by the QATS shared task– allows us to jointly address both aforementioned tasks (T1 and T2). Next, we propose robust systems based on using a freely available language learners corpus and a novel set of lexical complexity features (Section 4) to jointly address both tasks (T1 and T2) on a 1–5 level scale.

Finally, we: (1) show that our newly proposed lexical complexity features correlate well with human scores for sentence complexity (Section 5.1); (2) find the minimal feature set which leads to best classification performances (Section 5.2); and (3) show that our approach performs equally well even using a small-size (5,000-6,000 tokens) language learners corpus and could thus be adapted to other languages where such corpus exists or can be built (Section 5.3).

2 Related Work

Automatic assessment of sentence complexity has earlier been addressed as a pairwise ranking problem for the complex-simple English sentence pairs [Vajjala and Meurers, 2014; 2015; Ambati *et al.*, 2016]. The proposed systems relied on a

[1] Similar has already been shown for complex word identification task in lexical simplification systems [Paetzold and Specia, 2015].

[2] http://alt.qcri.org/semeval2016/task11/

great number of features, some of which depend on language-specific resources, such as psycholinguistic databases (e.g. the average age of acquisition, word familiarity rating, concreteness rating, etc.), or parsers (e.g. the number of subtrees and number of constituents per tree). This makes them: (1) difficult to adapt to different languages as many of the system components might not be available (or may lead to significantly lower performances) for other languages, and (2) difficult to adapt to automatic assessment of automatically simplified sentences (which are often ungrammatical and thus do not allow the use of parser-based features).

Another shortcoming of those studies is that they were only tested on the pairs of original and manually simplified sentences from the English Wikipedia – Simple English Wikipedia (EW–SEW) corpus [Vajjala and Meurers, 2014; Ambati *et al.*, 2016] and both the EW–SEW corpus and the OneStopEnglish corpus [Vajjala and Meurers, 2015]. This is a much easier task given that such sentence pairs usually represent strong paraphrases of each other,[3] as in the following example from the EW–SEW corpus:

EW: *In women, the larger mammary glands within the breast produce the milk.*

SEW: *The breast contains mammary glands.*

In text simplification, however, the differences between several possible versions of the same sentence are often much more subtle; the difference in output sentences of two ATS systems of similar architectures could be as little as using "*put* on the throne" instead of "*installed* on the throne" [Štajner *et al.*, 2016b]. Therefore, a system for automatic assessment of sentence complexity should ideally be sensitive even to such subtle differences. Furthermore, the automatically simplified sentences are often ungrammatical or have meaningless parts due to erroneous lexical substitutions. Both contribute to the sentence being perceived as more complex by the target user.

In contrast to the previously proposed features [Vajjala and Meurers, 2014; 2015; Ambati *et al.*, 2016], our (lexical) features have the advantage of requiring nothing but a language learners corpus with same texts adapted to various reading levels (and as will be shown later in Section 5.3, even a very small corpus suffices). This makes our systems for sentence complexity assessment – built upon those features – easily adaptable to other languages, for which such corpus exists (or can be built).

The winning QATS system [Štajner *et al.*, 2016a] uses the MT evaluation and MT quality estimation features, which are all computed by comparing the given simplified sentence with its original sentence, and are thus suitable only for evaluating ATS systems (the task in which we have the original/reference sentence) and cannot address the task of assessing complexity of original sentences. Our systems, in contrast, do not need the reference sentences, and thus can jointly address both tasks: complexity assessment of original sentences (T1), and complexity assessment of simplified sentences (T2).

[3]The EW–SEW dataset contains over 26% of strong paraphrases, and only 29% of sentence pairs with subtle differences between the original and simplified version [Štajner and Saggion, 2015].

3 Creation of Gold Standard Dataset

Given that human evaluation of sentence simplicity commonly uses a 5-level Likert scale, e.g. [Woodsend and Lapata, 2011; Saggion *et al.*, 2015], we decided not to follow the 3-level scale of the QATS shared task, but to compile a new dataset with a 5-level scale, enabling thus a more fine-grained sentence complexity classification.[4]

We randomly selected 150 sentences from various news stories[5] previously used in building an ATS system [Štajner and Glavaš, 2017], and 150 sentences from various English Wikipedia articles and automatically simplified them with five different ATS systems, two fully-fledged systems [Woodsend and Lapata, 2011; Angrosh *et al.*, 2014] and three lexical simplification systems [Paetzold and Specia, 2016; Glavaš and Štajner, 2015; Horn *et al.*, 2014]. Apart from including those 300 original sentences, we also included 150 manually simplified sentences from Simple English Wikipedia, which correspond to the 150 original Wikipedia sentences. Therefore, the dataset allows for jointly modelling sentence complexity of original, manually simplified, and automatically simplified sentences, building thus systems which can address both tasks (T1 and T2) mentioned in Section 1. After discarding repeated sentences (not all systems make changes to all original sentences), we had a total of 1131 sentences. Each sentence had 3–5 versions with subtle or more pronounced differences (see examples in Table 1).

Next, we asked three non-native speakers with high level of English (all studied at least 2 years in the UK) to evaluate "how easy is the sentence to understand" on a 1-5 scale, where 1 denotes very difficult and 5 very easy. Each annotator evaluated all 1131 sentences (we asked annotators never to annotate for longer than one hour without a pause in order to avoid the fatigue effect). We did not provide any guidelines as we did not want to bias their judgments towards any type of features. All variants of the same original sentence were presented one after another (in random order) to facilitate the annotation process by enabling implicit relative comparisons, and make the annotation as consistent as possible (in this way, two sentences with subtle differences are next to each other).

The average pairwise inter-annotator agreement (IAA), measured by quadratic Cohen's κ (to account for different levels of annotator disagreement as the task uses an ordinal scale), was 0.62 (0.58, 0.60, and 0.69, for each pair of annotators). After obtaining the satisfactory IAA, similar to those in previous TS evaluations [Štajner and Glavaš, 2017], we averaged the scores of three annotators and rounded them to the closest integer to obtain the 'gold standard' 1–5 complexity scores for each sentence (where 1 denotes a very complex, and 5 a very simple sentence). Several examples of sentences and their 'gold standard' complexity scores are given in Table 1. The distribution of sentences per score was the following: 217 scored 5, 309 scored 4, 283 scored 3, 210 score 2, and 112 scored 1.

[4]Data and code available at: http://web.informatik.uni-mannheim.de/sstajner/publications.

[5]http://takelab.fer.hr/data/evsimplify/

Sentence	Score
The school was founded in 1943 in the **ballots** of Westminster School in Little Dean's Yard, just behind Westminster Abbey.	3
The school was founded in 1943 in the **buildings** of Westminster School in Little Dean's Yard, just behind Westminster Abbey.	4
The school was founded in 1943 in the **precincts** of Westminster School in Little Dean's Yard, just behind Westminster Abbey.	3
Porcupines **are** the third **largest** of the rodents, **behind** the capybara and **the** beaver.	2
The porcupines **include** the third **biggest** rodent, **after** the capybara, and beaver**, and are not to be confused with hedgehogs**.	2
It was **eventually certified** gold by the RIAA in 2008.	4
It was **then awarded** gold by the RIAA in 2008.	5

Table 1: Examples of gold standard human scores for simplicity (differences between similar sentences are shown in bold).

Count	Level1	Level2	Level3	Level4	Level5
total unigrams	1,138,841	1,470,512	1,742,658	1,892,068	2,190,961
unique unigrams	28,904	36,361	45,441	51,914	55,892
unique bigrams	307,465	405,675	510,282	575,405	674,952
unique trigrams	702,715	934,767	1,152,266	1,277,984	1,495,966

Table 2: The number of unigrams and unique unigrams, bigrams and trigrams in the English Newsela corpus.

4 Approach

We use a freely available language learners corpus to learn lexical properties on different text complexity levels (Section 4.1). Based on those, we propose three simple methods for calculating phrase complexity level (Section 4.2) used for computing sentence complexity features (Section 4.3).

4.1 Corpora

We use the English part of the Newsela corpora[6] to learn lexical properties on different text complexity levels (the lists of unigrams, bigrams and trigrams occurring at each level, and their relative frequencies). The Newsela corpora contain 1913 original news articles in English and 244 in Spanish, simplified to four different learning levels by expert human editors. The applied manual simplifications are motivated by the Common Core Standards [Porter *et al.*, 2011] and the Lexile3 readability score [Xu *et al.*, 2015]. The total number of unigrams, and unique unigrams, bigrams, and trigrams on each English Newsela level are shown in Table 2.

We choose Newsela corpus over the EW–SEW corpus for two reasons: (1) the quality of simplifications in Newsela is controlled and guided by detailed standards, unlike the Simple Wikipedia whose quality is questioned by many researchers [Xu *et al.*, 2015]; and (2) Newsela offers five different complexity levels instead of only two levels offered by the EW–SEW corpus. Apart from Newsela corpus, two other corpora with controlled quality and five levels of simplifications are available for English, the WeeBit corpus [Vajjala and Meurers, 2014] aimed at native English speakers of different age levels, and the Cambridge English Exams corpora [Xia *et al.*, 2016] aimed at second language learners. However, those two corpora are not appropriate for learning lexical properties of different complexity levels as different levels do not contain the same stories. Therefore, the lexical properties learned from such corpora would not only be influenced by the text complexity, but also by the topics of the stories.

[6]https://newsela.com/data/

4.2 Phrase Complexity Level (PCL)

We calculate the phrase complexity level (PCL) for unigrams, bigrams, and trigrams by using three different methods:

1) **Relative Frequency (RF) method** relies only on the relative frequency of the given phrase in each of the five Newsela levels. The PCL is defined as the Newsela level at which the given phrase has the highest relative frequency.

2) **Lowest Level (LL) method** finds the first level (starting from the simplest one) on which the given phrase occurs. Once we find such a level, we assign it as the PCL and do not check if the phrase occurs on any higher level.

3) **Higher Levels (HL) method** defines the PCL as the Newsela level such that the given phrase occurs on that level and all other more complex levels, but it does not occur on the next simpler level.

In all three methods, we do not assign any PCL to the phrases which do not occur in any of the five text levels. Those phrases are not taken into account for calculating sentence complexity features.

It is important to note that the HL method differs from the LL method and is, actually, complementary to the LL method, as it often happens that some phrases do not appear on a certain level but they appear at the levels above and below. This is a common phenomenon in manual text simplification because: (1) human editors tend to perform significant content reduction during simplification [Saggion *et al.*, 2015]: (2) the lexical choices often depend on syntactic choices [Drndarevic *et al.*, 2012]; and (3) even the same guidelines lead to different lexical choices made by different human editors [Drndarevic *et al.*, 2012]. Therefore, low frequency phrases are sometimes omitted on certain levels. In the English Newsela corpus, for instance, there are 318 unigrams that appear both in Level 5 and in Level 3 but do not appear in Level 4. The unigram *healthiness* appears only in Levels 3 and 5, and not in any other. Therefore, it would be assigned the PCL 5 by the HL method and the PCL 3 by the LL method.

The assumption behind all three methods is that all four simplified levels contain simplifications of the same original texts (Level 5); otherwise the calculated PCL would also be influenced by the text topics and not only by their difficulty.

Code	Lexical complexity feature
max	the maximal PCL found
min	the minimal PCL found
avg	the average PCL found
count1	the number of phrases with level 1
count2	the number of phrases with level 2
count3	the number of phrases with level 3
count4	the number of phrases with level 4
count5	the number of phrases with level 5

Table 3: Lexical complexity features (calculated by each of the three methods and for each of the three phrase lengths).

Therefore, they should only be applied on the corpora which contain the same stories on each level.[7]

4.3 Sentence Complexity Features

For each sentence, we calculate eight lexical complexity features (see Table 3) for each phrase length (unigrams, bigrams, and trigrams) by using each of the three PCL methods. This results in a total of 72 lexical features to which we add the sentence length as a proxy for syntactic complexity. Note that for discriminating between complex and simple texts, sentence length has been shown to have a better discriminative power than syntactic complexity features based on the use of a parser [Štajner *et al.*, 2013], while at the same time being applicable on ungrammatical (automatically generated) sentences. While calculating lexical features, we exclude those phrases which do not appear in the whole English Newsela corpus. This helps in differentiating between grammatically correct and grammatically incorrect sentences, i.e. for grammatically incorrect sentences the number of found phrases at different PCL levels (the *count* features) will be lower than in similar but grammatically correct sentences.

5 Evaluation

We perform three sets of experiments (Sections 5.1– 5.3) to evaluate our approach.

5.1 Correlation with Human Scores

In order to assess the goodness-of-fit of our features to the sentence complexity assessment task, we first calculate the Pearson's correlation between all our features and the 'gold standard' human scores for this task (described in Section 3).

Out of 73 initial features (72 lexical features and the sentence length), 57 features showed significant correlation ($p < 0.01$) with the human scores for sentence complexity. The number of unigrams with the PCL 5 (*count5*) obtained using the RF method, and the sentence length, obtained the absolute Pearson's correlation score over 0.5 (Table 4).

The number of features significantly correlated (at a 0.01 level) with the human scores for simplicity, grouped by the phrase length, PCL method, and feature type, is presented in Table 5. Although none of the trigram-based features is among the seven features with the highest correlation to the human scores (Table 4), a great number of them still shows

[7]We also tried lemmatising all phrases and texts, but it did not lead to any significant improvements over the non-lemmatised versions.

Phrase	Feature	Method	Pearson
unigram	count5	RF	−0.523
	sentence length		−0.504
unigram	count1	LL	−0.499
unigram	count1	HL	−0.472
bigram	count5	RF	−0.430
bigram	count1	HL	−0.385
bigram	max	LL	−0.307

Table 4: Seven features with the highest Pearson's correlation (the absolute value over 0.300) with human scores for simplicity; all correlations significant at a 0.001 level.

Category	Type	#features
Phrase	unigram	18 (out of 24)
	bigram	20 (out of 24)
	trigram	19 (out of 24)
Method	Relative Frequency (RF)	21 (out of 24)
	Lowest Level (LL)	17 (out of 24)
	Higher Levels (HL)	19 (out of 24)
Feature	max	4 (out of 9)
	min	5 (out of 9)
	avg	4 (out of 9)
	count1	9 (out of 9)
	count2	8 (out of 9)
	count3	8 (out of 9)
	count4	9 (out of 9)
	count5	9 (out of 9)

Table 5: The number of features significantly correlated (at a 0.01 level) with the human scores for simplicity.

significant correlation with those. Out of the three PCL methods, the RF method produces the highest number of features significantly correlated with the human scores. Out of the eight feature types (Table 3), *count1*, *count4*, and *count5* show significant correlation with the human scores, regardless of the PCL method and the phrase length (Table 4).

5.2 Classification Experiments

In order to find out which subset of our 73 features lead to the best performing system on the sentence complexity task, we experimented with five classifiers and various feature sets.

Baselines. The best performing system (and most of the participating systems) of the QATS shared task cannot be applied on our task as they require reference sentences, which are not available in our, more general, sentence complexity task (where we also want to assess the complexity of original sentences). Also, as we also want to assess automatically generated (often ungrammatical) sentences, we do not try to apply previously proposed parser features to the pairwise ranking task of the EW–SEW sentence pairs [Vajjala and Meurers, 2014; 2015; Ambati *et al.*, 2016]. Therefore, as there are no existing systems addressing this task (the absolute sentence complexity assessment of original, manually simplified, and automatically simplified sentences, on a five-level scale), we build a strong baseline using three features which have shown the best performances in other similar tasks: sentence length (in words), average word length (in characters), and average word frequency in the whole Simple Wikipedia. The first two features are commonly used in various readability formulae (e.g. Flesch Reading Ease score [Flesch, 1949] or Fog

Feature set (# features)	Weighted F_1	ROC area	Accuracy	Quadratic κ	RMSE
Ours-correlated only (57)	0.65 ± 0.04	$\mathbf{0.87 \pm 0.02}$	66.60 ± 4.30	0.75 ± 0.01	$\mathbf{0.31 \pm 0.01}$
Ours-all (73)	0.65 ± 0.04	$\mathbf{0.87 \pm 0.02}$	64.92 ± 4.23	0.76 ± 0.01	$\mathbf{0.31 \pm 0.01}$
Ours-lexical only (72)	$\mathbf{0.66 \pm 0.04}$	$\mathbf{0.87 \pm 0.02}$	$\mathbf{65.67 \pm 4.08}$	$\mathbf{0.77 \pm 0.01}$	$\mathbf{0.31 \pm 0.01}$
Strong baseline	0.53 ± 0.04*	0.77 ± 0.03*	53.06 ± 4.41*	0.61 ± 0.01*	0.36 ± 0.01*
Majority class baseline	0.12 ± 0.00*	0.50 ± 0.00*	27.32 ± 0.27*	0.00 ± 0.00*	0.40 ± 0.01*

Table 6: The classification results (arithmetic mean with standard deviation) for our three feature sets, and the two baselines. The best results (by each evaluation metric) are presented in bold, and those which are significantly worse ($p < 0.01$, paired t-test) are presented with an '*'.

Experiments	Feature set (# features)	Weighted F_1	ROC area	Accuracy	Quadratic κ	RMSE
Feature type	Only count features: count1 ... count5 (45)	0.65 ± 0.04	$\mathbf{0.87 \pm 0.02}$	64.70 ± 4.26	0.76 ± 0.01	$\mathbf{0.31 \pm 0.01}$
	Only aggregate features: max, min, avg (27)	0.56 ± 0.04*	0.80 ± 0.03*	56.19 ± 4.12*	0.55 ± 0.01*	0.34 ± 0.01*
	Ours-lexical only (72)	$\mathbf{0.66 \pm 0.04}$	$\mathbf{0.87 \pm 0.02}$	$\mathbf{65.67 \pm 4.08}$	$\mathbf{0.77 \pm 0.01}$	$\mathbf{0.31 \pm 0.01}$
PCL method	RF (24)	0.62 ± 0.05*	0.86 ± 0.02 *	62.60 ± 4.67*	0.72 ± 0.01*	0.32 ± 0.01*
	LL (24)	0.61 ± 0.04*	0.84 ± 0.02*	60.59 ± 4.04*	0.66 ± 0.01*	0.32 ± 0.01*
	HL (24)	0.61 ± 0.04*	0.85 ± 0.02*	61.50 ± 4.03*	0.68 ± 0.01*	0.32 ± 0.01*
	RF + LL (48)	0.64 ± 0.04	$\mathbf{0.87 \pm 0.02}$	64.38 ± 4.41	0.72 ± 0.01*	$\mathbf{0.31 \pm 0.01}$
	RF + HL (48)	0.65 ± 0.04	$\mathbf{0.87 \pm 0.02}$	64.62 ± 4.21	0.76 ± 0.01	$\mathbf{0.31 \pm 0.01}$
	LL + HL (48)	0.64 ± 0.04*	0.86 ± 0.02*	63.86 ± 3.84	0.72 ± 0.01*	0.31 ± 0.01*
	RF + LL + HL (72)	$\mathbf{0.66 \pm 0.04}$	$\mathbf{0.87 \pm 0.02}$	$\mathbf{65.67 \pm 4.08}$	$\mathbf{0.77 \pm 0.01}$	$\mathbf{0.31 \pm 0.01}$
Phrase length	Unigrams only (24)	0.63 ± 0.04	0.86 ± 0.02	62.65 ± 4.08	0.77 ± 0.01	$\mathbf{0.31 \pm 0.01}$
	Bigrams only (24)	0.59 ± 0.04*	0.84 ± 0.02*	58.76 ± 4.29*	0.67 ± 0.01*	0.32 ± 0.01*
	Trigrams only (24)	0.46 ± 0.04*	0.74 ± 0.03*	46.07 ± 4.38*	0.48 ± 0.01*	0.37 ± 0.01*
	Unigrams+bigrams (48)	0.65 ± 0.04	$\mathbf{0.87 \pm 0.02}$	65.06 ± 4.33	$\mathbf{0.79 \pm 0.01}$	$\mathbf{0.31 \pm 0.01}$
	Unigrams+bigrams+trigrams (72)	$\mathbf{0.66 \pm 0.04}$	$\mathbf{0.87 \pm 0.02}$	$\mathbf{65.67 \pm 4.08}$	0.77 ± 0.01	$\mathbf{0.31 \pm 0.01}$

Table 7: The classification results (arithmetic mean with standard deviation) for features extracted using different feature types (*count* vs. *aggregate*), different methods for calculating PC, and features extracted using different phrase lengths. The best results are presented in bold, and those which are significantly worse ($p < 0.01$, paired t-test) than the best are presented with an '*'.

index [Gunning, 1952]) widely used for assessing text difficulty. Given that the readability formulae are applied on the text level and are not reliable on a sentence level, we just use their components and not the formulae themselves. The word frequency in the whole Simple Wikipedia has been used as a stand-alone feature in the winning system of the SemEval-2016 CWI task. Given that we assess the sentence complexity instead of the word complexity, we use this feature as the average word frequency in the Simple Wikipedia (averaged over all words in the given sentence).

Classifiers. We used five different classifiers: Logistic [le Cessie and van Houwelingen, 1992], SMOs – Weka implementation of SVM [Platt, 1998] with feature standardisation, JRip rule learner [Cohen, 1995], J48 – Weka implementation of C4.5 decision tree [Quinlan, 1993], and Random Forest [Breiman, 2001], in a 10-fold cross-validation setup with 10 repetitions in Weka Experimenter [Hall *et al.*, 2009]. The Random Forest classifier significantly outperformed all other four classifiers while preserving the same relative ranking of different systems. Therefore, we only report the performances of the Random Forest classifier, as the best one.

Evaluation. We evaluate all classifiers using the standard classification evaluation measures (weighted F_1 measure, ROC area, and accuracy). Additionally, to account for the fact that our classes are on an ordinal scale, we evaluate the classifiers using the quadratic Cohen's κ statistic (quadratic κ) and root mean squared error (RMSE).

Overall Results. Table 6 shows the results of the two baseline classifiers and our three systems, which use different sets of features: all 73 features (*Ours-all*), only lexical features (*Ours-lexical only*), and only those significantly correlated with human scores for simplicity (*Ours-correlated only*). All our systems significantly outperform both baselines. Our best system reaches a 0.66 weighted F-score, weighted ROC area of 0.87, 65.67% accuracy, a 0.77 quadratic κ, and a 0.31 RMSE, on average (the expert human annotators obtained a 0.62 quadratic κ on the same dataset (see Section 3)). The addition of the sentence length to the set of lexical features (our *count* features are not normalised with the sentence length and thus implicitly account for sentence length), or filtering out those features which do not show significant correlation with the human scores, do not significantly influence classification performances.

Influence of the Feature Type. It seems that the aggregate features (*min*, *max*, and *avg*) do not to contribute to the overall performance of the classifiers (Table 7), as similar classification performances can be obtained by using only the *count* features (*count1*, *count2*, *count3*, *count4*, and *count5*).

Influence of the PCL Method. It seems it is not necessary to use all three PCL methods (Table 7). The use of the RF method in combination with either of the other two methods (LL or HL) results in classification performances comparable to those of the system which uses all three PCL methods.

Influence of the Phrase Length. Using only *bigram* or *trigram* features significantly decreases classification performance (Table 7). More importantly, the use of only *unigram* features, or their combination with the *bigram* features, leads to almost equally good results as using all three phrase lengths.

Minimal Feature Set. Our previous sets of experiments showed that the combination of RF and HL methods lead to equally good results as the use of all three PCL methods, the use of only the *count* features leads to equally good results

Feature type	Weighted F_1	ROC area	Accuracy	Quadratic κ	RMSE
Unigram counts with RF+HL methods (10)	0.58 ± 0.04*	0.84 ± 0.02*	58.50 ± 4.39*	0.69 ± 0.01*	0.33 ± 0.01*
Unigram+bigram counts with RF+HL methods (20)	0.63 ± 0.05	$\mathbf{0.87 \pm 0.02}$	63.32 ± 4.63	0.73 ± 0.01	$\mathbf{0.31 \pm 0.01}$
Ours-all lexical features (72)	$\mathbf{0.66 \pm 0.04}$	$\mathbf{0.87 \pm 0.02}$	$\mathbf{65.67 \pm 4.08}$	$\mathbf{0.77 \pm 0.01}$	$\mathbf{0.31 \pm 0.01}$

Table 8: The classification results (arithmetic mean with standard deviation) for minimal feature set experiments. The best results are presented in bold, and those which are significantly worse ($p < 0.01$, paired t-test) than the best are presented with an '*'.

as the use of all eight features, and the use of only *unigram* features or their combination with *bigram* features leads to equally good results as the use of all three phrase lengths (Table 7). Therefore, we built two additional classifiers: (1) using only the *unigram count* features obtained by using only the RF and HL methods (10 features in total), and (2) using the combination of *unigram* and *bigram count* features obtained by using only the RF and HL methods (20 features in total). The first classifier performed significantly worse than the best system (all 72 lexical features), but the second system performed almost equally well as the best system.

5.3 Impact of the Corpus Size

In order to evaluate how easy it would be to adapt our system to other languages, we explored what is the minimal size of the corpus used to extract lexical properties on different levels, which leads to satisfactory performances on this task.

Using the bootstrapping resampling method with 100 repetitions, we extracted Newsela subcorpora in 16 different sizes: 5, 10, 20, 30, 40, 50, 100, 150, 200, 250, 500, 750, 1000, 1250, 1500, 1750 original texts (with their corresponding four simpler versions). We then performed 1,600 classification experiments on the 'gold standard' dataset, using the minimal feature set (unigram+bigram counts with RF+HL methods only) and obtaining the lexical properties each time from a different Newsela subcorpora. All experiments were performed in a 10-fold cross-validation setup using the Random Forest.

Surprisingly enough, we did not find any significant ($p < 0.01$, paired t-test) differences between the performances of the systems in which we extracted lexical properties from smaller portions of the Newsela corpus (not even the subcorpora which consist of only 5 original texts) and the performance of the system in which we extracted lexical properties from the whole Newsela corpus (1913 original articles). The smallest subcorpora (5 original texts) contained 4,657 tokens (1,238 unique tokens) in the original texts (Level 5), and 2,399 tokens (658 unique tokens) on the simplest level (Level 1), on average. These results show that we can obtain state-of-the-art results even if we extract lexical properties from a small corpus (4,000-5,000 tokens).

5.4 Our Approach on QATS shared task

One of the goals of the QATS shared task was to build 3-level classification systems for complexity assessment of automatically simplified sentences. We tested our system on the QATS shared task, by first applying the CfsSubsetEval feature selection algorithm [Hall and Smith, 1998] to the QATS training dataset (to select the best subset of our features) and then training the Random Forest classifier on the QATS training dataset using only the best subset of features. Our classifier obtained a 0.566 weighted F-score and 57.94% accuracy on the QATS test set, a performance comparable to the best ranked QATS system (0.564 weighted F-score and 57.14% accuracy).

6 Conclusions

Automatic sentence complexity assessment could bring many benefits to the fields of NLP and text accessibility. The systems should be able to assess original sentences (to identify those which need to be simplified) and manually and automatically simplified sentences (to choose those which are easy enough for the target reader). However, so far no systems have been proposed which would jointly address those tasks.

We built a 'gold standard' sentence complexity dataset containing original, manually simplified and automatically simplified sentences with human scores (on a 1–5 scale) assessing their absolute complexity. We then addressed the absolute sentence complexity as a five-level classification task and proposed novel lexical complexity features based on using nothing but a language learners corpus. The best classifier obtained a higher IAA (quadratic Cohen's κ of 0.77) with the gold standard marks than any pair of human annotators did among themselves (ranging from 0.58 to 0.69).

Most importantly, we showed that our methods perform well even if we use a very small language learners corpus to extract lexical properties (a 4,000-5,000 token corpus). This means that our approach could potentially be used on any language for which such a corpus exists, or can be built.

Acknowledgments

This work has been partially supported by the SFB 884 on the Political Economy of Reforms at the University of Mannheim (project C4), funded by the German Research Foundation (DFG).

References

[Aluísio and Gasperin, 2010] Sandra Maria Aluísio and Caroline Gasperin. Fostering Digital Inclusion and Accessibility: The PorSimples Project for Simplification of Portuguese Texts. In *Proceedings of YIWCALA*, pages 46–53, 2010.

[Ambati *et al.*, 2016] Bharat Ram Ambati, Siva Reddy, and Mark Steedman. Assessing Relative Sentence Complexity using an Incremental CCG Parser. In *Proceedings of NAACL-HLT*, pages 1051–1057, 2016.

[Angrosh *et al.*, 2014] Mandya Angrosh, Tadashi Nomoto, and Advaith Siddharthan. Lexico-syntactic text simplification and compression with typed dependencies. In *Procedings of COLING 2014*, pages 1996–2006, 2014.

[Breiman, 2001] Leo Breiman. Random Forests. *Machine Learning*, 45(1):5–32, 2001.

[Carroll *et al.*, 1999] John Carroll, Guido Minnen, Darren Pearce, Yvonne Canning, Siobhan Devlin, and John Tait. Simplifying text for language-impaired readers. In *Proceedings of EACL 1999*, pages 269–270, 1999.

[Cohen, 1995] William W. Cohen. Fast Effective Rule Induction. In *Proceedings of the ICML*, pages 115–123, 1995.

[De Belder and Moens, 2010] Jan De Belder and Marie-Francine Moens. Text simplification for children. In *Proceedings of the SIGIR workshop on accessible search systems*, pages 19–26, 2010.

[Drndarevic *et al.*, 2012] Biljana Drndarevic, Sanja Štajner, and Horacio Saggion. Reporting Simply: A Lexical Simplification Strategy for Enhancing Text Accessibility. In *Proceedings of Easy-to-read on the Web Symposium*, 2012.

[Flesch, 1949] Rudolf Flesch. *The art of readable writing*. Harper, New York, 1949.

[Glavaš and Štajner, 2015] Goran Glavaš and Sanja Štajner. Simplifying Lexical Simplification: Do We Need Simplified Corpora? In *Proceedings of ACL&IJCNLP*, pages 63–68, 2015.

[Gunning, 1952] Robert Gunning. *The technique of clear writing*. McGraw-Hill, New York, 1952.

[Hall and Smith, 1998] Mark A. Hall and Lloyd A. Smith. Practical feature subset selection for machine learning. In C. McDonald, editor, *Proceedings of the 21st Australasian Computer Science Conference (ACSC)*, pages 181–191. Berlin: Springer, 1998.

[Hall *et al.*, 2009] Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H. Witten. The weka data mining software: an update. *SIGKDD Explor. Newsl.*, 11:10–18, 2009.

[Horn *et al.*, 2014] Colby Horn, Cathryn Manduca, and David Kauchak. Learning a lexical simplifier using wikipedia. In *Proceedings of ACL 2014 (Short Papers)*, pages 458–463, 2014.

[le Cessie and van Houwelingen, 1992] Saskia le Cessie and Johannes C. van Houwelingen. Ridge Estimators in Logistic Regression. *Applied Statistics*, 41(1):191–201, 1992.

[Paetzold and Specia, 2015] Gustavo Henrique Paetzold and Lucia Specia. LEXenstein: A Framework for Lexical Simplification. In *Proceedings of ACL-IJCNLP 2015 System Demonstrations*, pages 85–90, 2015.

[Paetzold and Specia, 2016] Gustavo Henrique Paetzold and Lucia Specia. Unsupervised lexical simplification for non-native speakers. In *Proceedings of the AAAI*, 2016.

[Petersen and Ostendorf, 2007] Sarah E. Petersen and Mari Ostendorf. Text Simplification for Language Learners: A Corpus Analysis. In *Proceedings of the SLaTE Workshop*, pages 69–72, 2007.

[Platt, 1998] John C. Platt. Fast Training of Support Vector Machines using Sequential Minimal Optimization. In B. Schoelkopf, C. Burges, and A. Smola, editors, *Advances in Kernel Methods Support Vector Learning*, chapter 12, pages 41–65. MIT Press Cambridge, 1998.

[Porter *et al.*, 2011] Andrew Porter, Jennifer McMaken, Jun Hwang, and Rui Yang. Common Core Standards: The New U.S. Intended Curriculum. *Educational Researcher*, 40(3):103–116, 2011.

[Quinlan, 1993] Ross Quinlan. *C4.5: Programs for Machine Learning*. Morgan Kaufmann Publishers, CA, 1993.

[Saggion *et al.*, 2015] Horacio Saggion, Sanja Štajner, Stefan Bott, Simon Mille, Luz Rello, and Biljana Drndarevic. Making It Simplext: Implementation and Evaluation of a Text Simplification System for Spanish. *ACM Transactions on Accessible Computing*, 6(4):14:1–14:36, 2015.

[Vajjala and Meurers, 2014] Sowmya Vajjala and Detmar Meurers. Assessing the relative reading level of sentence pairs for text simplification. In *Proceedings of the EACL 2014*, pages 288–297, 2014.

[Vajjala and Meurers, 2015] Sowmya Vajjala and Detmar Meurers. Readability-based Sentence Ranking for Evaluating Text Simplification. Unpublished technical report, arXiv:1603.06009 [cs.CL], 2015.

[Štajner and Glavaš, 2017] Sanja Štajner and Goran Glavaš. Leveraging event-based semantics for automated text simplification. *Expert Systems with Applications*, 82:383 – 395, 2017.

[Štajner and Saggion, 2015] Sanja Štajner and Horacio Saggion. Translating from Original to Simplified Sentences using Moses: When does it Actually Work? In *Proceedings of RANLP 2015*, pages 611–617, 2015.

[Štajner *et al.*, 2013] Sanja Štajner, Biljana Drndarević, and Horacio Saggion. Corpus-based Sentence Deletion and Split Decisions for Spanish Text Simplification. *Computación y Systemas*, 17(2):251–262, 2013.

[Štajner *et al.*, 2016a] Sanja Štajner, Maja Popović, and Hanna Béchara. Quality Estimation for Text Simplification. In *Proceedings of the QATS Workshop*, pages 15–21, 2016.

[Štajner *et al.*, 2016b] Sanja. Štajner, Maja Popović, Horacio Saggion, Lucia Specia, and Mark Fishel. Shared Task on Quality Assessment for Text Simplification. In *Proceedings of the QATS Workshop*, pages 22–31, 2016.

[Woodsend and Lapata, 2011] Kristian Woodsend and Mirella Lapata. Learning to Simplify Sentences with Quasi-Synchronous Grammar and Integer Programming. In *Proceedings of the EMNLP*, pages 409–420, 2011.

[Xia *et al.*, 2016] Menglin Xia, Ekaterina Kochmar, and Ted Briscoe. Text Readability Assessment for Second Language Learners. In *Proceedings of the BEA Workshop*, pages 12–22, 2016.

[Xu *et al.*, 2015] Wei Xu, Chris Callison-Burch, and Courtney Napoles. Problems in Current Text Simplification Research: New Data Can Help. *TACL*, 3:283–297, 2015.