CATS: A Tool for Customized Alignment of Text Simplification Corpora

Sanja Štajner[1], Marc Franco-Salvador[2], Paolo Rosso[3], Simone Paolo Ponzetto[1]
[1] DWS Research Group, University of Mannheim, Germany
[2] Symanto Research, Nuremberg, Germany
[3] PRHLT Research Center, Universitat Politècnica de València, Spain
{sanja, simone} @informatik.uni-mannheim.de, marc.franco@symanto.net, prosso@prhlt.upv.es

Abstract

In text simplification (TS), parallel corpora consisting of original sentences and their manually simplified counterparts are very scarce and small in size, which impedes building supervised automated TS systems with sufficient coverage. Furthermore, the existing corpora usually do not distinguish sentence pairs which present full matches (both sentences contain the same information), and those that present only partial matches (the two sentences share the meaning only partially), thus not allowing for building customized automated TS systems which would separately model different simplification transformations. In this paper, we present our freely available, language-independent tool for sentence alignment from parallel/comparable TS resources (document-aligned resources), which additionally offers the possibility for filtering sentences depending on the level of their semantic overlap. We perform in-depth human evaluation of the tool's performance on English and Spanish corpora, and explore its capacities for classification of sentence pairs according to the simplification operation they model.

Keywords: text simplification, tools and resources, sentence similarity

1. Introduction

Automated text simplification (ATS) has the goal of automatically transforming sentence structure and lexical choices in a way that it provides better understanding and wider accessibility to large audiences. The main obstacle for successful supervised ATS is the scarcity and limited size of parallel TS corpora which would contain original sentences and their manual simplifications. The parallel TS corpus for Brazilian Portuguese, compiled for the purposes of the *PorSimples* project (Aluísio et al., 2008) contains around 4,500 aligned sentences, and the parallel TS corpus for Spanish, compiled for the purposes of the *Simplext* project (Saggion et al., 2015) contains only around 1,000 aligned sentences. The largest existing TS comparable corpora is the English Wikipedia – Simple English Wikipedia (EW–SEW), consisting of 170,000 sentence pairs (Kauchak, 2013), or 150,000 full matches and 130,000 partial matches in the newer version (Hwang et al., 2015). In both cases, the sentences were automatically aligned from comparable English Wikipedia and Simple English Wikipedia articles. However, the use of EW–SEW dataset for modeling TS has been disputed (Amancio and Specia, 2014; Štajner et al., 2015; Xu et al., 2015) for several reasons: (1) the simplified articles are not necessarily direct simplifications of the original articles; (2) the quality of simplifications is not checked; (3) the dataset does not cover sentence splitting which is one of the most common operations in text simplification.

The Newsela corpora[1] of document-aligned news texts, manually simplified at four different simplification levels have been freely available for a few years for research purposes. These corpora have several advantages over the EW–SEW dataset (Xu et al., 2015; Štajner et al., 2017): (1) simplified texts present direct simplifications of the original articles; (2) simplification was performed by trained human editors, following strict guidelines; (3) by sentence-aligning those corpora one can get training material for simplifications at various levels, i.e. train different simplification models depending on the intended reader group; and (4) they provide comparable training material in two languages, English and Spanish. At the beginning of 2016, the Newsela corpora contained around 2,000 original news articles in English and around 250 original news articles in Spanish (both with their corresponding manually simplified versions at four different simplification levels).

The current state-of-the-art systems for automatic sentence-alignment of original and manually simplified text are the Greedy Structural WikNet (GSWN) method (Hwang et al., 2015) used for sentence-alignment of original and simple English Wikipedia, and the HMM-based (using Hidden Markov Model and Viterbi algorithm) method (Bott and Saggion, 2011) used for sentence-alignment of the Spanish Simplext corpus (Saggion et al., 2015). The HMM-based method can be applied to any language as it does not require any language-specific resources. It is based on two hypothesis: (H1) that the original order of information is preserved, and (H2) that every 'simple' sentence has a corresponding 'original' sentence. The GSWN method does not assume H1 or H2, but it only allows for '1-1' sentence alignments (which is very restricting for TS) and it is language-dependent as it requires the English Wiktionary[2].

In this paper, we present a freely available tool for sentence- and paragraph-alignment from document-aligned TS corpora, the CATS (Customized Alignment for Text Simplification) tool.[3] The tool offers two main functionalities:

1. **CATS-Align**: sentence- or paragraph- alignment of parallel texts; and

[1] `https://newsela.com/`

[2] `https://www.wiktionary.org/`

[3] The CATS tool and documentation can be downloaded from: `https://github.com/neosyon/SimpTextAlign`

2. **CATS-Measure**: three different sentence (or paragraph) similarity measures which can be further used to filter retrieved sentence/paragraph pairs for customised modeling of text simplification operations.

The CATS-Align has two main advantages over the state-of-the-art GSWN method:

- CATS is language-independent and resource light. Two out of three similarity metrics that CATS require only pre-trained word embeddings in the language for which is to be used, while the third similarity metric does not even require word embeddings (it is based on character n-gram matching).
- The GSWN method only allows for '1-1' sentence alignments, while CATS-aligns additionally offers a possibility for building a dataset which covers sentence splitting ('1-n' alignments).

Similarly to the HMM-based method, our alignment methods assume the hypothesis H2. We provide them in both variants, using the hypothesis H1 and without it (Section 2.2). The detailed human evaluation of our methods and the HMM-based method (Section 3) for both English and Spanish showed that our methods are significantly better, especially when aligning sentences from distant complexity levels.

The CATS tool was released together with our previous paper (Štajner et al., 2017). In that work, we were purely interested in augmenting the parallel datasets for training ATS systems and we performed an intrinsic evaluation of the tool only on the English part of Newsela corpora, and an extrinsic evaluation by using newly aligned dataset in a PBSMT approach to English ATS.

Here we build on our previous work by describing the full potential of the CATS tool: (1) for automatic sentence-alignment of both English and Spanish corpora; and (2) for automatic filtering of aligned sentence pairs according to the simplification operation: content deletion, information addition, and paraphrasing without significant semantic change.

In Section 2 we present different modes of the CATS tool. In Section 3, we present a detailed human evaluation and error analysis of the CATS-Align on both English and Spanish Newsela corpora. We evaluate CATS-Measure automatically on the 'gold standard' English Wikipedia dataset for classification of sentence pairs in three classes (*full matches*, *partial matches*, and *no match*) in Section 4. Finally, in Section 5, using the labels assigned by human annotators on the Newsela dataset during the human evaluation of CATS-Align, we test whether the CATS-Measure can be used to automatically classify the automatically-aligned sentence pairs (aligned by CATS-Align) into four classes depending on the type of simplification operation they model.

2. The CATS

Our CATS software can work in two different regimes:

- **CATS-Measure** for providing three different similarity measures that can be applied either on paragraph or sentence level;
- **CATS-Align** for choosing best paragraph or sentence alignments in a given document-aligned corpus.

2.1. CATS-Measure

CATS-Measure provides similarity measures for three different sentence/paragraph similarity methods:

1. **C3G:** The Character N-Gram (CNG) (Mcnamee and Mayfield, 2004) similarity model (with *n* set to 3) with log TF-IDF weighting (Salton and McGill, 1986).
2. **WAVG:** The continuous skip-gram model (Mikolov et al., 2013b) of the TensorFlow toolkit[4] on the English Wikipedia. For each text snippet (i.e. sentence or paragraph, depending on the task) we average its word vectors to obtain a single representation of its content. This setting has shown good results in other NLP tasks (e.g. for selecting out-of-the-list words (Mikolov et al., 2013a), or for language variety identification (Franco-Salvador et al., 2015)).
3. **CWASA:** The Continuous Word Alignment-based Similarity Analysis (CWASA) model (Franco-Salvador et al., 2016) was initially proposed for plagiarism detection with excellent results. Unlike the WAVG method, CWASA does not average word vectors and was thus proposed as more adequate for long texts.

In all three methods, similarity between the vectors is calculated using the cosine similarity. In WAVG and CWASA methods we use 300-dimensional vectors calculated with the continuous skip-gram model. We use the September 2016 Wikipedia dumps as input to train the English vectors. The model uses negative sampling, context windows of size 10, and 20 negative words for each sample. For Spanish, we use the freely available pretrained 300-dimensional vectors obtained using the skip-gram model with negative sampling on a large collection of various Spanish corpora with a total of approximately 1.5 billion words (Cardellino, 2016).

2.2. CATS-Align

For aligning sentences or paragraphs from a document-aligned TS corpora, CATS-Align offers two different alignment strategies (MST and MST-LIS) depending on whether we assume the hypothesis H1 (see Section 1) that the simplified text presents the information in the same order as the original text:

- **Most Similar Text (MST):** Having a set of 'simple' text snippets S, a set of 'complex' text snippets C, and one of the similarity methods (Section 2.1), MST compares similarity scores of all possible pairs (s_i, c_j), and aligns each $s_i \in S$ with the closest one in C.
- **MST with Longest Increasing Sequence (MST-LIS):** MST-LIS uses the hypothesis H1. It first uses the MST strategy, and then postprocess the output by extracting – from all obtained alignments – only those alignments $l_i \in L$, which contain the longest increasing sequence of offsets j_k in C.

[4] https://www.tensorflow.org/

Ex.	Original	Simplified
1a	Hand parts take up to 10 hours to print and another couple of hours to assemble with elastic cords to keep the hands open.	Hand pieces take up to 10 hours to print.
1b	Hand parts take up to 10 hours to print and another couple of hours to assemble with elastic cords to keep the hands open.	Putting them together takes another couple of hours.
2a	With one his wife bought him for Father's Day, sheets of colored plastic, and free designs and advice found online, he made a hand for about $20	Chi made his first 3-D hand with a printer his wife bought him for Father ś Day.
2b	With one his wife bought him for Father's Day, sheets of colored plastic, and free designs and advice found online, he made a hand for about $20	He found free designs online.
2c	With one his wife bought him for Father's Day, sheets of colored plastic, and free designs and advice found online, he made a hand for about $20	With them, he printed a hand for about $20.

Table 1: Examples of '1-n' alignments obtained by CATS-Align on the Newsela corpora (Newsela, 2016), which can be used either for modeling sentence compression (each example separately) or sentence splitting (by merging examples 1a–1b, and merging examples 2a–2c).

In order to allow for '1–n' alignments (i.e. sentence splitting), we allow for repeated offsets of C in L.

The 'simple' text snippets not contained in L are included in the set U of unaligned snippets.

Finally, we align each $u_m \in U$ by restricting the search space in C to those offsets of 'complex' text snippets that correspond to the previous and the next aligned 'simple' snippets. For instance, if $L = \{(s_1, c_4), (s_3, c_7)\}$ and $U = \{s_2\}$, then the search space for the alignments of s_2 is reduced to $\{c_4 ... c_7\}$.

We denote the MST-LIS alignment strategy by adding '*' to the name of the similarity method (e.g. C3G*).

2.2.1. Modeling Sentence Splitting and Compression

In both alignment strategies (MST and MST-LIS), we allow the same original sentence to be aligned with multiple simple sentences, in order to allow for modeling both sentence splitting and sentence compression later on. This is one of the important differences between our alignment models and the state-of-the-art GSWN method, which only allows '1-1' alignment and thus does not offer a possibility for building a dataset which covers sentence splitting ('1-n' alignments). An example of our customisable alignment tool is presented in Table 1. While each of the separate examples (1a–2c) can be later used for modeling sentence compression, by merging the examples 1a–1b and 2a–2c, we also build good training materials for modeling sentence splitting operations.

2.2.2. Two-Step Alignment

Additionally, the CATS-Align offers two-step alignment option, by first performing paragraph-alignment, and then sentence-alignment within each pair of aligned paragraphs. In this option, paragraphs and sentences can be aligned by any of the six previously mentioned strategies (three similarity methods times two alignment strategies), and not necessarily the same one. Two-step C3G alignment (C3G-2s) has shown best results in the extrinsic evaluation when used for building ATS systems (Štajner et al., 2017).

3. Human Evaluation on Newsela Datasets

We randomly selected 10 original English articles and 10 original Spanish articles, together the four corresponding simpler versions (at different levels of simplification) for each of them, and sentence-aligned them with seven different alignment strategies offered by the CATS tool: C3G, C3G*, CWASA, CWASA*, WAVG, WAVG*, C3G-2step, and the HMM-based alignment tool (Bott et al., 2012). Then we asked two native speakers of English (first trained on additional 3 original articles and their corresponding simplified versions) and two native speakers of Spanish (first trained in the same manner) to classify the obtained sentence pairs (a total of approx. 3,500 sentence-pairs for each language) in one of the four classes:

- 3: full match (full semantic overlap),
- 2: partial match (partial semantic overlap where the original sentence contains less information than the simplified sentence)
- 1: partial match (partial semantic overlap where the original sentence contains more information than the simplified sentence),
- 0: no match (no semantic overlap).

While sentence pairs with full matches can be used to model paraphrasing, sentence pairs with partial matches can be used to model deletions (class '1' where the original sentence contains more information than the simplified sentence) or additions (class '2' where the original sentence contains less information than the simplified sentence). Several examples from different classes are presented in Table 2. As can be seen, the CATS tool can successfully align sentences with full semantic overlap which differ only by one lexical/phrasal substitution (the second and the third example in Table 2), as well as those which are much stronger paraphrases of each other (the first example in Table 2). It can also align the sentences which have only partial semantic overlap (examples 4a, 4b, and 5 in Table 2).

3.1. Results of Human Evaluation

The results of this human evaluation are presented in Tables 3 and 4, as a percentages of different classes/matches. Given the human effort needed for such evaluation, we focused only on three level pairs: aligning the sentences from

Ex.	Original	Simplified	Class	Sim.
1	**After focusing on** the latest **artificial** limb technology**, he began to hunt** for **more basic** options.	**He looked into** the latest **prosthetic** limb technology **and began hunting** for **cheaper, less complicated** options.	3	0.52
2	Like many researchers, **entrepreneurs** and even artists in recent years, he turned to the 3-D printer.	Like many researchers, **businesspeople** and even artists in recent years, he turned to the 3-D printer.	3	0.80
3	**With** one his wife bought him for Father's Day, sheets of colored plastic, and free designs and advice found online, he made a hand for about $20.	**By using** one his wife bought him for Father's Day, sheets of colored plastic, and free designs and advice found online, he made a hand for about $20.	3	0.95
4a	**A non-profit** group called Women On 20s**, formed to convince President Barack Obama to put a woman's image on the $20 note, already has done some polling.**	**There is a** group called Women On 20s.	1	0.45
4b	**A non-profit group called Women On 20s, formed to convince President Barack Obama to put** a woman's **image** on the $20 **note, already has done some polling.**	**It wanted** a woman's **picture** on the $20 **bill**.	1	0.26
5	The plastic comes out in layers.	**They melt and** the plastic comes out in layers.	2	0.82
6	**Lew has said that Hamilton's image will remain part of** the new $10 bill.	The new $10 bill **will have the picture of a woman, he said**.	0	0.47
7	**These are some of the candidates to be** the first woman on **U.S. currency notes** in more than a **century**.	**Who will be** the first woman on **American money** in more than a **100 years**?	0	0.30

Table 2: Examples of different classes of alignments obtained by CATS-Align on the Newsela corpora (Newsela, 2016), together with their similarity scores obtained by C3G-2s alignment strategy. Differences between original and simplified sentences are presented in bold.

the original articles (Level 0) and the first simpler level (Level 1), aligning the sentences for the original articles (Level 0) and the simplest articles (Level 4), and aligning the sentences from the two simplest levels (Level 3 and Level 4). Due to the nature of simplification operations needed to be applied between levels 0 and 1, and those needed between levels 0 and 4, we expect a greater lexical and n-gram overlap between the sentences needed to align between levels 0 and 1, than those sentences needed to align between levels 0 and 4. Furthermore, we are interested in exploring whether the success rate of the alignment tool stays stable whenever we align two neighbouring levels, thus taking into account both 0–1 and 3–4 alignments. Finally, we investigate whether the success rate stays stable across the two languages.

3.1.1. Comparison with the State of the Art

In both languages and on all level pairs, the CATS alignments were able to find higher number of full and partial matches than the state-of-the-art HMM alignment method (Tables 3 and 4). The differences in the percentage of full matches found by the CATS alignments and the HMM method are particularly pronounced when we align 0-1 levels (up to 9.4% difference on the English dataset, and up to 10.8% difference on the Spanish dataset). The differences in the percentages of partial matches modeling deletion (*Part-Del*) between the CATS alignments and the HMM method while aligning 0-1 levels are noticeable on the Spanish dataset (up to 13.4% difference), while there is no much difference on the English dataset (only up to 3.3% difference). In aligning other level pairs (3-4 and 0-4) the differences in the percentages of partial matches modeling deletion were significant regardless the language.

3.1.2. The Influence of Hypothesis H1

We noticed that the use of hypothesis H1 reduces the percentage of full matches regardless of language, level pairs, and similarity measure. However, it sometimes increases the number of partial matches which model deletion (see Tables 3 and 4).

3.1.3. CATS Alignments across Languages

When comparing the performances of CATS alignments across the two languages, we find that alignment of 0-1 levels yields in slightly higher percentage of full matches on the English dataset than on the Spanish dataset, but at the cost of having lower percentage of partial matches modeling deletion (*Part-Del*). When aligning Level 3 with Level 4, we find similar percentage of full matches in both languages and a higher number of partial matches (deletions) on the English dataset. In the case of aligning Level 0 with Level 4, we also have a higher percentage of full matches on the English dataset than on the Spanish dataset, in addition to a higher percentage of partial matches. However, the differences in the percentage of full matches and partial matches between the two languages might not reflect the performances of the system on those languages but rather the nature of simplifications performed on the Newsela arti-

Align	Method	Class			
		Full	Part-Add	Part-Del	No
0-1	C3G	**69.6%**	2.8%	23.8%	**3.9%**
0-1	C3G*	67.4%	2.8%	**24.3%**	5.5%
0-1	CWASA	69.1%	2.8%	**24.3%**	**3.9%**
0-1	CWASA*	66.9%	2.8%	**24.3%**	6.1%
0-1	WAVG	**69.6%**	2.2%	**24.3%**	**3.9%**
0-1	WAVG*	67.4%	2.2%	**24.3%**	6.1%
0-1	HMM	60.2%	2.8%	21.0%	16.0%
3-4	C3G	44.5%	0.9%	**32.7%**	**21.8%**
3-4	C3G*	42.7%	0.9%	33.2%	23.2%
3-4	CWASA	**45.5%**	1.4%	31.3%	**21.8%**
3-4	CWASA*	42.7%	0.9%	31.3%	25.1%
3-4	WAVG	44.1%	1.4%	32.2%	22.3%
3-4	WAVG*	41.2%	0.9%	**32.7%**	25.1%
3-4	HMM	38.9%	0.9%	23.7%	36.5%
0-4	C3G	**10.0%**	0.5%	43.6%	46.0%
0-4	C3G*	6.2%	0.0%	**48.8%**	**45.0%**
0-4	CWASA	9.5%	0.5%	33.6%	56.4%
0-4	CWASA*	6.6%	0.0%	35.1%	58.3%
0-4	WAVG	9.5%	0.5%	45.0%	**45.0%**
0-4	WAVG*	6.6%	0.0%	43.6%	49.8%
0-4	HMM	4.7%	0.0%	20.1%	74.4%

Table 3: Distribution of different sentence-alignments according to the human evaluation on the English Newsela corpora. For each pair of levels, the highest percentage of *full* and *part-del* matches, and the lowest percentage of *no* matches are shown in bold.

Align	Method	Class			
		3 (Full)	2 (Add)	1 (Del)	0 (No)
0-1	C3G	61.4%	3.4%	**31.8%**	**3.4%**
0-1	C3G*	55.7%	3.4%	30.7%	10.2%
0-1	CWASA	61.4%	3.4%	29.5%	5.7%
0-1	CWASA*	55.7%	3.4%	30.7%	10.2%
0-1	WAVG	**62.5%**	2.3%	28.4%	6.8%
0-1	WAVG*	55.7%	2.3%	**31.8%**	10.2%
0-1	HMM	50.6%	1.1%	18.4%	29.9%
3-4	C3G	**44.8%**	1.0%	**26.7%**	**27.6%**
3-4	C3G*	41.9%	1.0%	25.7%	31.4%
3-4	CWASA	43.8%	1.9%	21.0%	33.3%
3-4	CWASA*	41.9%	1.9%	24.8%	31.4%
3-4	WAVG	41.9%	1.9%	21.9%	34.3%
3-4	WAVG*	41.0%	1.9%	23.8%	33.3%
3-4	HMM	38.1%	0.0%	18.1%	43.8%
0-4	C3G	**3.8%**	0.0%	37.1%	59.0%
0-4	C3G*	1.9%	0.0%	**43.8%**	**54.3%**
0-4	CWASA	**3.8%**	1.0%	35.2%	60.0%
0-4	CWASA*	1.0%	0.0%	35.2%	63.8%
0-4	WAVG	**3.8%**	1.0%	35.2%	60.0%
0-4	WAVG*	2.9%	0.0%	26.7%	70.5%
0-4	HMM	1.0%	0.0%	10.5%	88.6%

Table 4: Distribution of different sentence-alignments according to the human evaluation on the Spanish Newsela corpora.

cles in those two languages, i.e. it is possible that the simplification in Spanish leads to more sentence splittings thus reflected in higher number of partial matches, while the simplification in English leads to more paraphrasing without sentence splitting and thus more full matches. The performances of the CATS-Align are rather reflected in number of *no matches*, which according to the results in Tables 3 and 4 indicate similar performances of CATS-Align for both languages in the case of aligning 0-1 levels, and a slightly better performances of the tool on the English than on the Spanish dataset in the case of 3-4 and 0-4 alignments.

3.2. Error Analysis

We found that over 50% of original sentences from Level 0 get split into two or more (up to even five sentences) simple sentences when simplifying into Level 4. This sometimes results in low scores (*no match*) by human evaluation of isolated sentence pairs, although all simple sentences that correspond to the same original sentence, when seen together, perfectly match the original sentence (Table 5). In other words, when treated separately, some of the sentences aligned between levels 0 and 4 present false negatives, as they will later be merged together for training ATS systems, where they will then represent good training material for sentence splitting. These type of errors could be avoided by performing human evaluation on already merged sentences (those that together model sentence splitting), but in that case we would not have the count of correct deletions and additions, which might be useful for tasks other than simplification, or for modeling these specific text simplification operations.

4. Automatic Evaluation on Wikipedia

The 'gold standard' Wikipedia dataset for sentence-alignment (Hwang et al., 2015) contains pairs of sentences and their 'gold label' (*full match*, *partial match*, or *no match*) with the same meaning as in our manual evaluation task. As the C3G, CWASA and WAVG methods output similarity score for each sentence pair, we use them to predict the labels for the sentence pairs in the Wikipedia dataset. Table 6 shows the F_1-measures obtained by our systems (using different combinations of sentence similarity measures and classification algorithms) and the state-of-the-art GSWN system, as well as several baselines used by Hwang *et al.* (2015) on two classification tasks: classifying between *Good&GoodPartial* matches vs. *Others* (Task 1), and between *Good* matches vs. *Others* (Task 2). The HMM-method requires full texts to use the H1 hypothesis and thus cannot be successfully applied to these tasks.

Detailed results of our various classification methods on Tasks 1 and 2, presenting precision (P), recall (R) and F_1-measure (F) on the *Good & GoodPartial* class (Task1) or the *Good* class (Task 2), are presented in Table 7. We observe similar behaviour of our similarity metrics in both tasks; the C3G method obtains significantly better recall than the word-embedding-based methods (CWASA and WAVG), while CWASA and WAVG obtain better precision than the C3G. All three methods combined together significantly boost both recall and F_1-measure, significantly outperforming all previously proposed methods on Task 1, and

Ex.	Simplified (Level 4)	Original (Level 0)
1	Todos los estudiantes tienen que hacer el examen SAT para ser admitidos en la universidad.	En otras palabras, si los estudiantes afroamericanos con las puntuaciones del SAT más bajas que las de sus compañeros blancos también reciben peores calificaciones en la universidad, o se cambian con más frecuencia a carreras "más fáciles", entonces, para empezar, seguramente estaban menos preparados para la universidad.
2	Algunas personas opinan que los estudiantes de color están menos preparados para entrar a la universidad.	En otras palabras, si los estudiantes afroamericanos con las puntuaciones del SAT más bajas que las de sus compañeros blancos también reciben peores calificaciones en la universidad, o se cambian con más frecuencia a carreras "más fáciles", entonces, para empezar, seguramente estaban menos preparados para la universidad.
3	Dicen esto cuando los estudiantes afroamericanos tienen calificaciones más bajas que las de sus compañeros blancos en ese examen.	En otras palabras, si los estudiantes afroamericanos con las puntuaciones del SAT más bajas que las de sus compañeros blancos también reciben peores calificaciones en la universidad, o se cambian con más frecuencia a carreras "más fáciles", entonces, para empezar, seguramente estaban menos preparados para la universidad.
4	También, cuando reciben calificaciones más bajas en la universidad o se cambian a carreras "menos difíciles".	En otras palabras, si los estudiantes afroamericanos con las puntuaciones del SAT más bajas que las de sus compañeros blancos también reciben peores calificaciones en la universidad, o se cambian con más frecuencia a carreras "más fáciles", entonces, para empezar, seguramente estaban menos preparados para la universidad.

Table 5: An example of sentence pairs which obtained score 0 (no match) each, but in fact present a good example of an '1–n' alignment obtained by CATS-Align C3G method on the Newsela corpora (Newsela, 2016).

Approach	*Task1*	*Task2*
C3G+CWASA+WAVG	**.643**	.705
C3G+CWASA	.621	.680
C3G+WAVG	.602	.691
CWASA+WAVG	.506	.664
C3G	.612	.695
CWASA	.490	.671
WAVG	.481	.650
GSWN (Hwang et al., 2015)	.607	**.712**
Unconst.WordNet (Hwang et al., 2015)	.515	.636
Ordered Vec.Space (Hwang et al., 2015)	.415	.564
Unconstr. Vec.Space (Hwang et al., 2015)	.431	.550

Table 6: F_1-measures on Task1 (*Good & Good Partial vs. Others*) and Task2 (*Good vs. Others*). The best results for each task are shown in bold.

CATS measures	Task 1			Task 2		
	P	R	F	P	R	F
C3G+CWASA+WAVG	.808	**.534**	**.643**	.829	**.614**	**.705**
C3G+CWASA	.780	.516	.621	.780	.603	.680
C3G+WAVG	.760	.498	.602	.791	**.614**	.691
CWASA+WAVG	**.829**	.364	.506	.808	.563	.664
C3G	.777	.505	.612	.803	.574	.669
CWASA	.792	.355	.490	.827	.502	.625
WAVG	.791	.346	.481	**.830**	.527	.645

Table 7: Detailed results on Task 1 (*Good & Good Partial vs. Others*) and Task2 (*Good vs. Others*). The best results for each measure are shown in bold.

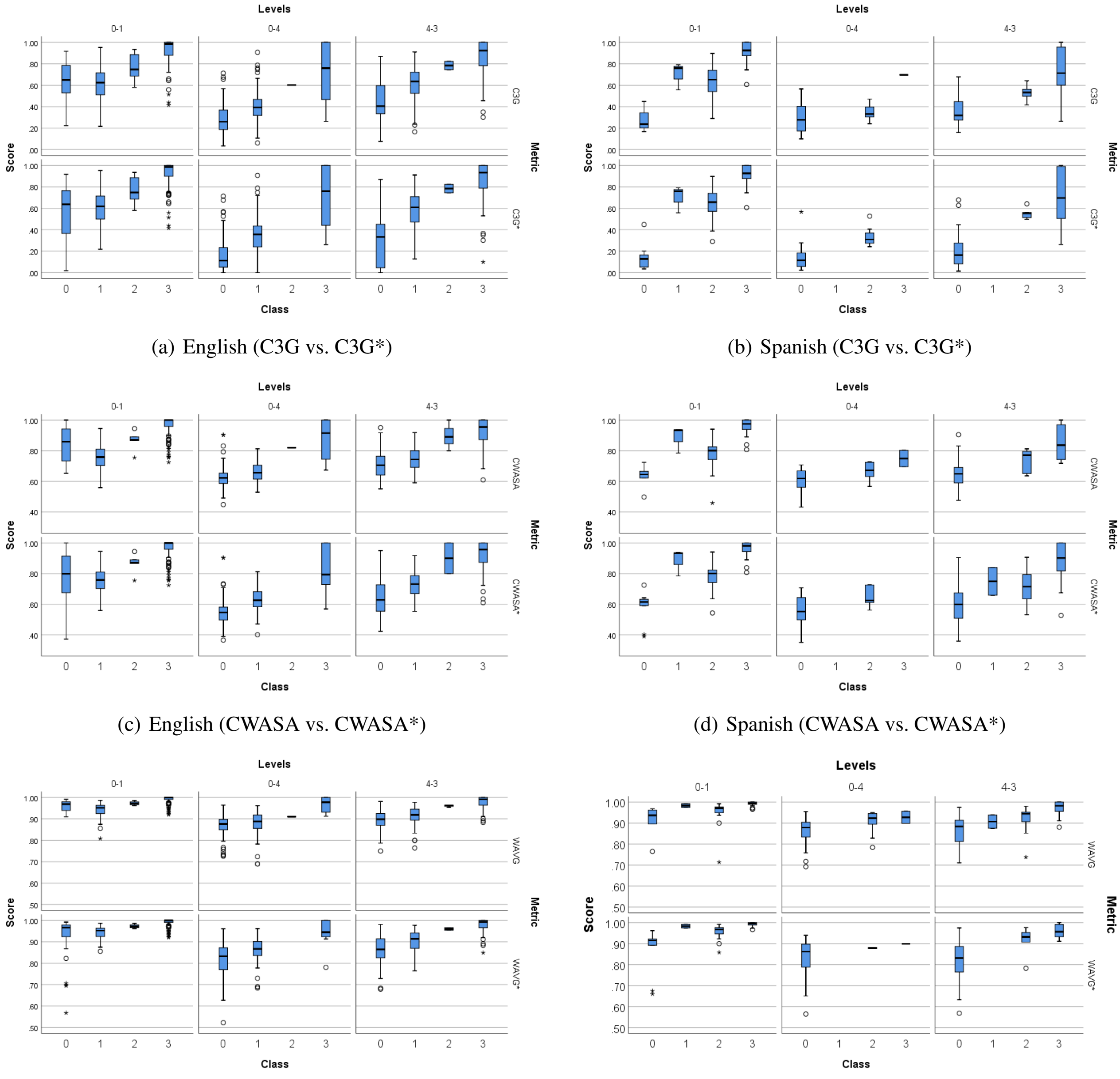
obtaining comparable results to the state of the art on Task 2 (see Table 6). Here is important to mention that, unlike the current state-of-the-art method on Task 2 (Hwang et al., 2015), our methods are **language-independent**, **resource-light**, and allow for retrieving material for sentence splitting (**allow for '1–n' matches**).

5. Automatic Classification of Alignments

Finally, we explore if the similarity measures provided by CATS-Measure can be used to classify the aligned sentence pairs according to the type of simplification operation they model.

5.1. Distribution of Similarity Scores

We first explore the distribution of similarity scores across different transformation types and different text levels on both English and Spanish human annotated datasets (Figure 1). As can be seen, all six sentence similarity metrics (C3G, C3G*, CWASA, CWASA*, WAVG, WAVG*) seem to have better discriminatory power between the four classes (0–3) on the Spanish dataset than on the English dataset (discriminatory power seen as the overlap of boxplots for different classes), with the C3G and C3G* being the best among the six metrics. Here is important to mention that the metrics do not have to be able to distinguish between the classes 1 (insertion) and 2 (deletion). To discriminate between those two classes (in the case of similar metrics scores) we can use the difference in the sentence length between the two sentences, i.e. deletions and insertions should lead to the opposite sign when we subtract the length of the original sentence (in words) from the length of the simplified sentence (in words). The distributions of sentence similarity scores (Figure 1) indicate that we can expect better classification results for Spanish than for English experiments.



(a) English (C3G vs. C3G*)

(b) Spanish (C3G vs. C3G*)

(c) English (CWASA vs. CWASA*)

(d) Spanish (CWASA vs. CWASA*)

(e) English (WAVG vs. WAVG*)

(f) Spanish (WAVG vs. WAVG*)

Figure 1: Distribution of the six similarity scores across different classes, text levels, and languages.

5.2. Classification Experiments

For classification experiments, we use the labels assigned by human evaluators as the ‘gold standard’ labels. As features, we use the similarity metrics obtained by CATS-Measure (one metric at the time) and the differences in word count between the original and simplified sentence (in order to distinguish between *additions* and *deletions*) which achieve similar scores for semantic similarity by all our similarity metrics (see Figure 1). Given that we observed certain differences in distribution of similarity metrics across different text levels, we also experiment with adding the level pair as an additional (third) feature for the classification.

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).

As can be seen from the classification results presented in Table 8 (only for the best classifier, logistic), although CATS-Align achieved lower performances on the Spanish dataset, the *no matches* can easier be automatically filtered for Spanish than for English (lower number of false positives for Spanish than for English). The percentage of false positives for the *no match* class (fPos) indicate that if we are interested only in filtering out *no matches*, we can successfully achieve this by training the classifiers on a small number of human annotated sentence pairs, especially for Spanish (where, in the best scenario, using the WAVG*, difference in sentence length, and the level pair, the percent-

Features	English		Spanish	
	w-F	fPos	w-F	fPos
C3G+len+levels	71.5	32.5%	82.6	16.7%
C3G+len	74.9	32.0%	82.3	21.4%
C3G*+len+levels	**75.3**	32.2%	84.1	16.3%
C3G*+len	**75.3**	32.2%	**84.8**	16.3%
CWASA+len+levels	70.1	27.5%	80.9	19.1%
CWASA+len	70.2	30.2%	83.8	14.9%
CWASA*+len+levels	72.6	**26.6%**	80.5	15.3%
CWASA*+len	72.3	27.0%	81.4	13.9%
WAVG+len+levels	65.0	45.5%	77.0	23.1%
WAVG+len	65.1	52.5%	78.6	30.3%
WAVG*+len+levels	69.7	42.0%	82.9	**13.6%**
WAVG*+len	69.6	42.0%	83.1	15.3%

Table 8: Results of automatic classification of sentence pairs into four categories (*no match*, *deletions*, *additions*, and *full matches*) presented as the weighted average F_1 measure (*w-F*) and the percentage of false positives for the *no match* class (*fPos*), i.e. cases in which *no match* was classified as any other category. The best scores achieved for each classification evaluation metric (*w-F* and *fPos*), and for each language, are presented in bold.

age of false positives for the *no match* class is only 13.6). If we are interested in classifying sentence pairs by different transformation operations, this can again be successfully achieved with classifiers trained on the small number of instances, with better results for Spanish than for English (weighted F-measure of 84.8 for Spanish, and 75.3 for English).

The results in Table 8 also indicate that specifying the level pair from which sentences were aligned improves the F-measure on the Spanish classification task (though not necessarily decreases the number of false positives for the *no match* class), but has no effect on the English classification task.

6. Conclusions

One of the main problems of the state-of-the-art automatic text simplification systems is the absence and the small size of parallel datasets (pairs of original sentences and their manually simplified versions) which leads to insufficient coverage of supervised systems. The CATS tool presented in this paper offers several different ways of sentence- and paragraph-aligning of document-aligned texts on different text complexity levels. It additionally offers three sentence similarity metrics which can be applied on sentence pairs and used for automatically classifying simplification operations as *full matches*, *additions*, *deletions*, and *no matches*. Our detailed human evaluation of the alignment module (CATS-Align) showed that it can successfully align sentence pairs from document-aligned corpora in English and Spanish. The results of classification experiments confirmed that the sentence similarity measures offered by our CATS-Measure can be used as features for classification of sentence pairs as *full matches*, *additions*, *deletions*, and *no matches* on both English and Spanish Newsela corpora. More importantly, they showed that wrongly aligned sentence pairs can be automatically filtered out by classifiers built on small size human annotated datasets (approximately 1,000 instances).

Finally, the resource-light and language-independent sentence similarity metrics offered by CATS-Measure performed similar to the state-of-the-art systems for classifying sentence pairs from English Wikipedia and Simple English Wikipedia as *full matches*, *partial matches*, and *no matches*, proving thus that they are not effective only on the news domain but also on the encyclopedic domain.

The CATS tool with both CATS-Align and CATS-Measure options is freely available on: `https://github.com/neosyon/SimpTextAlign`.

7. Acknowledgements

We would like to thank Newsela for making their data available, and to Luke Orland for his valuable comments and suggestions. Part of the work presented in this article was performed during second author's research visit to the University of Mannheim, supported by Contact Fellowship awarded by the DAAD scholarship program "STIBET Doktoranden". The work of Paolo Rosso has been partially funded by the SomEMBED TIN2015-71147-C2-1-P MINECO research project.

8. Bibliographical References

Aluísio, S. M., Specia, L., Pardo, T. A., Maziero, E. G., and Fortes, R. P. (2008). 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.

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.

Bott, S. and Saggion, H. (2011). An Unsupervised Alignment Algorithm for Text Simplification Corpus Construction. In *Proceedings of the Workshop on Monolingual Text-To-Text Generation*, pages 20–26. ACL.

Bott, S., Saggion, H., and Figueroa, D. (2012). A Hybrid System for Spanish Text Simplification. In *Proceedings of the Workshop on Speech and Language Processing for Assistive Technologies (SLPAT), held in conjunction with NAACL-HLT 2012*, pages 75–84.

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

Cardellino, C. (2016). Spanish Billion Words Corpus and Embeddings. http://crscardellino.me/SBWCE/.

Cohen, W. W. (1995). Fast Effective Rule Induction. In *Proceedings of the Twelfth International Conference on Machine Learning*, pages 115–123.

Franco-Salvador, M., Rangel, F., Rosso, P., Taulé, M., and Martí, M. A. (2015). Language variety identification using distributed representations of words and documents. In *Proceeding of the 6th International Conference of CLEF on Experimental IR meets Multilinguality, Multimodality, and Interaction (CLEF 2015)*, volume LNCS(9283), pages 28–40. Springer-Verlag.

Franco-Salvador, M., Gupta, P., Rosso, P., and Banchs, R. E. (2016). Cross-language plagiarism detection over continuous-space- and knowledge graph-based representations of language. *Knowledge-Based Systems*, 111:87 – 99.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., and Witten, I. H. (2009). The weka data mining software: an update. *SIGKDD Explor. Newsl.*, 11:10–18.

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.

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

Mcnamee, P. and Mayfield, J. (2004). Character n-gram tokenization for European language text retrieval. *Information Retrieval*, 7(1):73–97.

Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013a). Efficient estimation of word representations in vector space. In *Proceedings of Workshop at International Conference on Learning Representations*.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J. (2013b). Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems 26*, pages 3111–3119.

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

Platt, J. C. (1998). Fast Training of Support Vector Machines using Sequential Minimal Optimization. In B. Schoelkopf, et al., editors, *Advances in Kernel Methods 6 Support Vector Learning*, chapter 12, pages 41–65. MIT Press Cambridge.

Quinlan, R. (1993). *C4.5: Programs for Machine Learning*. Morgan Kaufmann Publishers, San Mateo, CA.

Saggion, H., Štajner, S., Bott, S., Mille, S., Rello, L., and Drndarevic, B. (2015). Making It Simplext: Implementation and Evaluation of a Text Simplification System for Spanish. *ACM Transactions on Accessible Computing*, 6(4):14:1–14:36.

Salton, G. and McGill, M. J. (1986). *Introduction to Modern Information Retrieval*. McGraw-Hill, Inc.

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.

Xu, W., Callison-Burch, C., and Napoles, C. (2015). Problems in Current Text Simplification Research: New Data Can Help. *Transactions of the Association for Computational Linguistics (TACL)*, 3:283–297.