Complex Word Identification Using Character n-grams

Maja Popović
Humboldt University of Berlin
Berlin, Germany
maja.popovic@hu-berlin.de

Abstract

This paper investigates the use of character n-gram frequencies for identifying complex words in English, German and Spanish texts. The approach is based on the assumption that complex words are likely to contain different character sequences than simple words. The multinomial Naive Bayes classifier was used with n-grams of different lengths as features, and the best results were obtained for the combination of 2-grams and 4-grams. This variant was submitted to the Complex Word Identification Shared Task 2018 for all texts and achieved F-scores between 70% and 83%. The system was ranked in the middle range for all English tracks, as third of fourteen submissions for German, and as tenth of seventeen submissions for Spanish. The method is not very convenient for the cross-language task, achieving only 59% on the French text.

1 Introduction

Complex Word Identification (CWI) refers to identification of words which are considered by readers from a specific target audience to be complex. The CWI task is the first step towards the lexical simplification task which aims at improving the readability of texts: a lexical simplification system should replace the identified complex words with their simpler synonyms. Some of these systems have a CWI module at the beginning of their pipeline, e.g. (Paetzold and Specia, 2015) whereas some perform the CWI task implicitly, such as (Glavaš and Štajner, 2015).

The first shared task on CWI was organized at the SemEval 2016 (Paetzold and Specia, 2016) where 21 teams submitted 42 systems trained to predict whether words in a given context were complex for a non-native English speaker. Following the success of the first CWI shared task and additional findings reported in (Zampieri et al., 2017), the second shared task has been organised at the BEA workshop 2018 (Yimam et al., 2018) featuring a multilingual dataset. The dataset consists of training and testing sets for three languages: English, German and Spanish, as well as French test set for cross-lingual CWI. The goal was to predict which words could be difficult for a non-native speaker, based on annotations collected from a mixture of native and non-native speakers. The predictions could be submitted in the form of class labels (complex or simple) and/or in the form of complexity probabilities.

This work proposes the use of character n-grams for complex word identification. The main assumption is that complex words contain different character sequences than simple words, i.e. that the combination of particular characters is related to the complexity of a word. Additional motivation is the successful use of character n-grams for machine translation evaluation metrics in recent years (Stanojević and Sima'an, 2014; Popović, 2015; Wang et al., 2016). The results of Machine Translation Metrics Shared Tasks[1] (Bojar et al., 2017) have shown that these metrics correlate very well with human judgments for all analysed target languages, which indicates that character sequences carry some important information.

We used the multinomial Naive Bayes classifier, although the assumption about independence between different n-grams was certainly not valid. The motivation to conduct our first experiments with character n-grams using this classifier was its often use as a baseline for text classification (McCallum and Nigam, 1998; Kibriya et al., 2004; Lohar et al., 2017). Since Naive Bayes probabilities are generally not reliable, we participated only in the binary classification task.

[1] http://www.statmt.org/wmt17/metrics-task.html

Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications, pages 341–348
New Orleans, Louisiana, June 5, 2018.

Although the relation between character n-grams and word complexity intuitively depends on the language, we still decided to investigate cross-lingual CWI and to participate in this track.

1.1 Related work

Several different techniques for identifying complex words were investigated by (Shardlow, 2013) which include word frequency, word length and syllable counts among others, but no character sequences.

The first CWI shared task (Paetzold and Specia, 2016) featured 42 systems based on different techniques and using different features such as semantic, morphological, lexical, as well as word frequencies which are reported to be a very important factor for CWI.

One of the submitted systems (Mukherjee et al., 2016) used Naive Bayes classifier with morphological, semantic and lexical features, however no character n-grams were investigated.

Another system (Zampieri et al., 2016) used probabilities of word character trigrams and sentence character trigrams together with word length and sentence length to measure orthographic difficulty. These features together with the word frequency features are used for three classifiers: Random Forest, Nearest neighbour and SVM. Nevertheless, no results regarding the contribution of character trigram features were reported.

Number of vowels, number of syllables and number of characters (word length) together with word frequencies in corpora were investigated in (Yimam et al., 2017b), but no experiments with character n-grams were conducted.

2 Character n-grams and multinomial Naive Bayes classifier

For each labelled word, all character n-grams of given length(s) and their frequencies were extracted and the word was represented as a "bag of n-grams". Decision on which n-gram length(s) to concentrate is far from trivial since, to our best knowledge, no similar experiments have been conducted before. Therefore, we started with individual n-gram lengths from 2 to 6, following the findings from machine translation metric task where lenghts above 6 did not bring any improvements. Our preliminary experiments showed that introducing six-grams degraded the performance so we kept the lengths up to 5. As for mixed lengths, the best preliminary results were obtained for 2-gram, 3-gram and 4-gram combinations, so we concentrated on these variants.

Table 1 presents two complex and three simple English words with their 2-grams, 3-grams and 4-grams and corresponding frequencies. Under the (very) naive assumption of conditional independence between individual n-grams, these frequencies are then used for estimating the class-contidition probabilities of the Naive Bayes multinomial model:

$$\hat{c} = \arg\max_{c} P(c) \prod_{i=1}^{N_{ngr}} P(ngr_i \mid c) \qquad (1)$$

where $P(ngr_i|c)$ is the conditional probability that the n-gram ngr_i occurs in a word with the class value c, and N_{ngr} is the total number of distinct n-grams, i.e. the dimension of the feature vector. $P(c)$ is the prior probability that a word has class label c.

For the multinomial model, these two probabilities can be estimated as relative frequencies in the following way:

$$\hat{P}(ngr_i \mid c) = \frac{count(ngr_i, c)}{\sum_{i=1}^{Ngr} count(ngr_i, c)} \qquad (2)$$

where the numerator represents the number of occurences of the n-gram ngr_i in a word with class label c, and the denominator represents the number of occurences of all n-grams in this class. The smoothing probability for unseen n-grams was set to 0.001.

The prior class probability can be estimated as:

$$\hat{P}(c) = \frac{count(c)}{count(words)} \qquad (3)$$

where $count(c)$ represents the number of words with class label c and $count(words)$ represents the total number of labelled words.

If the words in Table 1 and their 4-grams were used for training, the prior class probabilities for simple ("S") and complex ("C") words would be $P(S) = 3/5 = 0.60$ and $P(C) = 2/5 = 0.4$. Class condition probabilities for the 4-gram "frug" would be $P(frug|S) = 0$ and $P(frug|C) = 1/5 = 0.2$, and for the 4-gram "real" $P(real|S) = 0.25$, $P(real|C) = 0$. The 4-ram "lity" would have similar probabilities for the complex and for

word	*"bag of n-grams": 2-grams, 3-grams, 4-grams and their frequencies*	*class*
frugality	fr:1 ru:1 ug:1 ga:1 al:1 li:1 it:1 ty:1 fru:1 rug:1 uga:1 ali:1 lit:1 ity:1 frug:1 ruga:1 ugal:1 gali:1 lity:1	C
reefs	re:1 ee:1 ef:1 fs:1 ree:1 eef:1 efs:1 reef:1 eefs:1	C
banana	ba:1 an:2 na:2 ban:1 ana:2 nan:1 bana:1 anan:1 nana:1	S
coral	co:1 or:1 ra:1 al:1 cor:1 ora:1 ral:1 cora:1 oral:1	S
reality	re:1 ea:1 al:1 li:1 it:1 ty:1 rea:1 eal:1 ali:1 lit:1 ity:1 real:1 eali:1 alit:1 lity:1	S

Table 1: Examples of two complex and three simple words with their 2-grams, 3-grams and 4-grams and corresponding frequencies.

the simple class since it appears both in "frugality" and in "reality": $P(lity|S) = 1/4 = 0.25$, $P(lity|C) = 1/5 = 0.20$.

3 Data

The organisers of the shared CWI task provided all participants with training and test data sets for English, German and Spanish. For French, only test data set was provided since it was intended for the cross-lingual CWI task. The English data set consists of mixture of professionally written news (News), non-professionally written news (WikiNews), and Wikipedia articles (Wiki). German, Spanish and French data sets contain data taken from German, Spanish and French Wikipedia pages. Data statistics is presented in Table 2.

Each sentence in the English data set was annotated by 20 people, 10 native and 10 non-native speakers. Each sentence in the German, Spanish and French data sets was annotated by 10 people, a mixture of native and non-native speakers. Annotators were provided with the surrounding context of each sentence, i.e. a paragraph, then asked to mark words they think would be difficult to understand for children, non-native speakers, and people with language disabilities. Annotators were enabled not only to annotate individual words, but also several consecutive words as complex. The details about the data sets can be found in (Yimam et al., 2017b) and (Yimam et al., 2017a).

4 Results

As mentioned in Section 2, the main part of our experiments was to determine which n-gram lenghts to include in the classifier. Preliminary experiments showed that the individual lengths of 2,3,4 and 5 should be further investigated, as well as combinations of 2- and 4-grams, 3- and 4-grams, as well as 2-, 3- and 4-grams.

All these variants were investigated for three scenarios: (i) standard classification, where each training set corresponds to the development set, (ii) classification with the extended English training corpus, where all English training corpora were concatenated and used for classifying each of the development sets, and (iii) cross-lingual classification, where training sets of other two languages were used for each language.

The comparison of the methods was carried out on the development sets in terms of complex word F-score and overall accuracy.

4.1 Standard set-up

In the standard set-up, each development set was classified using its corresponding training set, both in terms of domain and of language. Table 3 represents the obtained resuls, with best F-scores / accuracies in bold.

It can be noted that the combination of 2-grams and 4-grams is the best option for allmost all texts. It is second ranked (and very close to the best one) only for the accuracy of English news. As for the individual n-grams, the best performance is obtained by 4-grams. The scores are improving when increasing n-gram length up to 4, and then drop for 5-grams (except for the accuracy of English News and German Wikipedia). It can also be seen that in general, combining different n-gram lenghts works better than using the individual ones.

4.2 Concatenated English training corpus

Since the English data set contained three domains: Wikipedia, News and WikiNews, the question about effects of enlarging the training set arised: will the use of a larger training corpus from

#words	English			German	Spanish	French
domain	Wiki	News	WN	Wiki	Wiki	Wiki
Train	5551	14002	7746	6151	13750	0
Dev	694	1764	870	795	1622	0
Test	870	2095	1287	959	2233	2251

Table 2: Data statistics: the number of instances for each training, development and test set used in the CWI 2018 shared task.

n-gram	English			German	Spanish
length(s)	Wiki	News	WikiNews	Wiki	Wiki
2	64.7 / 64.8	63.2 / 70.5	61.8 / 67.6	60.7 / 69.6	55.9 / 68.2
3	67.5 / 68.7	72.6 / 77.8	64.8 / 71.5	62.5 / 68.6	62.0 / 70.3
4	67.5 / 69.3	75.9 / 81.2	68.9 / 75.4	60.9 / 69.9	63.4 / 73.1
5	61.1 / 67.3	75.0 / **81.7**	64.5 / 74.6	57.3 / 70.2	58.2 / 72.6
24	**69.9 / 70.9**	**76.7 / 81.3**	**69.9 / 75.7**	**65.3 / 72.6**	**64.7 / 73.6**
34	68.3 / 69.2	75.9 / 80.4	68.5 / 74.0	62.2 / 69.2	**64.7** / 72.4
234	68.4 / 69.4	75.4 / 79.6	**69.9** / 75.0	62.9 / 69.4	**64.4** / 72.1

Table 3: F-score for complex word class / accuracy for English, German and Spanish development sets.

different domains lead to better results or not? In order to answer this question, each of the three English development sets was also classified using the concatenated English training corpus containing all three domains and the results are presented in Table 4. These results show that enlarging the training corpus generally helps.

The smallest improvements can be observed for the News text, probably because the News training corpus is the largest one, as can be noted in Table 2. Another finding is that for the larger training set, individual 3-grams, 4-grams and 5-grams can outperform the n-gram combinations. A possible explanation is that the reliability of longer character sequences is increased when a larger training corpus with more instances is used. When the three n-gram length combinations are compared on the larger training set, "24" still outperforms the other two except for the Wikipedia set.

4.3 Cross-lingual classification

In order to explore cross-language classification, each of the Wikipedia development sets was classified using the training corpora of another two languages. English News and WikiNews development sets were not used in order to avoid possible effects of domain mixing. The results in Table 5 show that the method is, as mentioned in Section 1, indeed not very appropriate for cross-lingual classification since the character combinations are generally language dependent – the drop in F-score and accuracy is large, in the range of 10 to 15 absolute points.

As for the n-gram lengths, combination "24" is useful, although mostly for English. For German and Spanish, 3-grams and 5-grams outperformed the n-gram combinations. As for the usage of different languages, no advantage of one "foreign" language over another was observed – the best results are rather similar for both "external" languages. For example, the F-score for English is slightly better when the German training set is used, and accuracy is slightly better when the system was trained on the Spanish text. The fact that none of the language pairs is closely related might have an important influence on these results.

4.4 Confusion analysis

The results described in previous sections have shown the following:

- combination of 2-grams and 4-grams is the best option for the standard setting, and performs decently also for enlarged English training corpus as well as for cross-lingual classification;
- individual 3-grams, 4-grams and 5-grams outperform the combinations when a larger English corpus is used.

n-gram length(s)	Wiki dev		News dev		WikiNews dev	
	Wiki train	all train	News train	all train	WN train	all train
2	64.7 / 64.8	61.2 / 63.6	63.2 / 70.5	61.9 / 69.8	61.8 / 67.6	63.0 / 68.8
3	67.5 / 68.7	68.6 / 69.7	72.6 / 77.8	71.5 / 77.5	64.8 / 71.5	**73.4** / 74.0
4	67.5 / 69.3	**73.5 / 74.6**	75.9 / 81.2	76.0 / 81.3	68.9 / 75.4	**73.4** / 78.4
5	61.1 / 67.3	66.8 / 71.8	75.0 / **81.7**	75.9 / **82.4**	64.5 / 75.6	71.2 / **79.0**
24	**69.9 / 70.9**	70.9 / 72.0	**76.7** / 81.3	**76.4** / 81.3	**69.9 / 75.7**	73.3 / 77.9
34	68.3 / 69.2	73.3 / 74.1	75.9 / 80.4	76.2 / 81.0	68.5 / 74.0	72.8 / 77.2
234	68.4 / 69.4	71.4 / 72.2	75.4 / 79.6	75.4 / 80.2	**69.9** / 75.0	72.5 / 77.0

Table 4: F-score for English complex word class / accurracy for domain-specific and concatenated training set.

n-gram length(s)	English development		German development		Spanish development	
	es-train	de-train	en-train	es-train	en-train	de-train
2	50.7 / 59.4	60.1 / 59.5	48.9 / 54.3	49.9 / 62.6	55.4 / 55.7	53.9 / 57.4
3	58.0 / 60.5	60.4 / 58.6	**55.6** / 55.0	49.6 / 56.5	**55.8** / 54.2	55.0 / 58.6
4	57.3 / 62.5	51.7 / 57.2	53.8 / 61.2	55.6 / **64.0**	51.8 / 58.4	45.8 / 59.7
5	41.7 / 59.6	38.2 / 57.5	34.9 / **62.5**	33.3 / 63.1	38.2 / **61.0**	24.4 / **61.8**
24	58.9 / **63.0**	**61.4 / 61.4**	53.3 / 56.7	**57.0** / 63.9	53.4 / 54.4	52.5 / 57.2
34	**59.7** / 61.7	59.6 / 57.1	53.7 / 53.0	53.9 / 58.0	54.6 / 53.1	54.7 / 57.1
234	59.3 / 61.4	61.1 / 58.2	51.6 / 51.6	56.0 / 60.9	54.8 / 53.4	**55.4** / 56.7

Table 5: Cross-language classification: F-score for complex word class / accurracy for cross-language classification.

In order to better understand the above findings, confusion analysis was carried out for all n-gram lengths and for all Wikipedia development sets in all three set-ups.

Table 6 shows the percentages of (non-)confusions: C-C and S-S represent correctly classified instances, C-S stands for complex words classified as simple, and S-C for simple words classified as complex. The results show the following:

- 5-grams are very good in identifying simple words: less than 10% of them are classified as complex. Nevertheless, they are absolutely the worse in labelling complex words: for German and Spanish texts, they even label more complex instances incorrectly than correctly (red numbers).

- the combination "24" is very good in labelling complex words, although often outperformed by one of the other two combinations; the percentages in the majority of those cases are very close, though.

- the same combination, "24", is the best of all three combinations for labelling simple words, although clearly outperformed by 5-grams and 4-grams.

The described findings indicate that the combination "24", despite not always yielding the best scores, is the most balanced and the most stable one over all set-ups. Therefore, this variant was submitted for all shared task tracks.

It should be noted that the confusions were also analysed for the cross-lingual classification showing the very same behaviour for 5-grams and for the "24" variant. As for other n-gram lenghts, a number of different large confusion percentages was observed, indicating once again that the method is not convenient for cross-lingual CWI.

5 Official shared task results

Following all the findings described in previous sections, we decided to submit the "24" variant, i.e. the combination of 2-grams and 4-grams, to all shared task tracks. For each of the three English test sets, we sent two submissions: one classified using the corresponding in-domain training corpus, and one classified using the concatenated training corpus. For the French test set, we sent four submissions: one classified using English

(a) English – in-domain training corpora

n-gram order(s)	Wiki				News				WikiNews			
	C-C	C-S	S-C	S-S	C-C	C-S	S-C	S-S	C-C	C-S	S-C	S-S
2	32.3	16.6	18.6	32.6	25.3	14.1	15.4	45.2	26.2	14.4	18.0	41.4
3	32.4	16.4	14.8	36.3	29.5	10.0	12.2	48.3	26.2	14.4	14.1	45.3
4	31.8	17.0	13.7	37.5	29.6	9.8	9.0	51.5	27.2	13.3	11.3	48.2
5	25.6	23.2	**9.5**	**41.7**	27.4	12.0	**6.2**	**54.3**	23.1	17.4	**7.9**	**51.5**
24	**33.9**	**15.0**	14.1	37.0	30.7	8.8	9.9	50.7	28.2	12.4	11.8	47.6
34	33.3	15.6	15.3	35.9	30.8	8.6	11.0	49.6	28.3	12.3	12.7	45.7
234	33.1	15.7	14.8	36.3	**31.3**	**8.2**	12.2	48.3	**29.0**	**11.6**	13.3	46.1

(b) English – concatenated training corpus

n-gram order(s)	Wiki				News				WikiNews			
	C-C	C-S	S-C	S-S	C-C	C-S	S-C	S-S	C-C	C-S	S-C	S-S
2	28.7	20.2	16.1	35.0	24.5	14.9	15.3	45.2	26.6	14.0	17.1	42.3
3	33.1	15.7	14.6	36.6	28.7	10.8	12.1	48.5	28.0	12.5	13.4	46.0
4	35.2	13.7	11.7	39.5	29.5	10.0	8.7	51.9	29.9	10.7	10.9	48.5
5	28.4	20.5	**7.8**	**43.4**	28.2	11.2	**6.6**	**53.9**	26.0	14.6	**6.4**	**53.0**
24	34.0	14.8	13.1	38.0	30.3	9.1	9.6	51.0	30.2	10.3	11.7	47.7
34	**36.0**	**13.2**	12.7	38.5	**30.4**	**9.0**	10.0	50.6	**30.4**	**10.1**	12.6	46.8
234	34.7	14.1	13.7	37.5	30.3	9.1	10.7	49.9	30.3	10.2	12.8	46.7

(c) German

n-gram order(s)	C-C	C-S	S-C	S-S
2	23.5	18.5	11.9	46.1
3	26.2	15.8	15.6	42.4
4	23.4	18.6	11.4	46.5
5	**20.0**	**22.0**	**8.8**	**50.2**
24	25.8	16.2	11.2	46.8
34	25.4	16.6	14.2	43.8
234	**25.9**	**16.1**	14.5	43.5

(d) Spanish

n-gram order(s)	C-C	C-S	S-C	S-S
2	20.2	20.1	11.6	48.1
3	24.2	16.1	13.6	46.2
4	23.3	17.0	10.0	49.7
5	**19.0**	**21.2**	**6.2**	**53.6**
24	24.2	16.0	10.4	49.4
34	**25.3**	**14.9**	12.7	47.0
234	25.2	15.0	12.9	46.9

Table 6: Confusion analysis for the English, German and Spanish development sets: C-C and S-S are correctly classified complex and simple words, C-S stands for complex words classified as simple, and S-C for simple words classified as complex.

Wikipedia training corpus, one classified using the concatenated English training corpus, one classified using the Spanish training corpus, and one using the German training corpus. For the German and Spanish test sets, one submission was sent for each.

The official accurracies for the best system, for all our submissions and for the worst system are shown in Table 7 together with the ranks (in parenthesis).

All our monolingual submissions were ranked in the middle, some better than others. The best rank is achieved for German (3 of 14) and the worst for Spanish (10 from 17). The obtained accurracies are all above 70%, the German being the lowest one and the English News the highest one. For the cross-lingual task, our submissions were ranked very low, with one of the submissions being the worst one. However, it should be noted that the use of the Spanish training set yielded the best result: this indicates that the method could potentially be used for closely related languages, however this should be further examined in future work.

All the results indicate that there is a potential for using character n-grams for complex word identification, however more experiments should be carried out and several refinements should be applied.

6 Summary and outlook

In this paper, we have proposed the use of character n-grams for complex word idenfitication starting from the assumption that character sequences in complex words are often different than those in simple words. We carried out extensive experiments with multinomial Naive Bayes classifier with n-grams of different lengths as features, and found out that using 2-grams and 4-grams is the most stable option in this configuration. Our system was ranked in a middle-range position for all tracks except for the cross-lingual track where it was ranked very low – this was not surprising since frequencies of character sequences in words are intuitively rather language-dependent. Our official accurracy scores range from 70% to 83% for English, German and Spanish texts and from 50% to 59% for French cross-lingually classified text.

Our experiments described in this work together with the official shared task results indicate that the use of character n-grams for complex word identification has a potential, but the methods should be further investigated and improved. First of all, other classifiers without independency assumption should be investigated. In addition, using context (surrounding words and their n-grams) should be investigated as well.

References

Ondřej Bojar, Yvette Graham, and Amir Kamran. 2017. Results of the WMT17 Metrics Shared Task. In *Proceedings of the Second Conference on Machine Translation (WMT 17)*, pages 489–513, Copenhagen, Denmark.

Goran Glavaš and Sanja Štajner. 2015. Simplifying lexical simplification: do we need simplified corpora? In *Proceedings of 53rd annual meeting of the Association for Computational Linguistics and 7th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing*, pages 63–68, Beijing, China.

Ashraf M. Kibriya, Eibe Frank, Bernhard Pfahringer, and Geoffrey Holmes. 2004. Multinomial Naive Bayes for Text Categorization Revisited. In *Proceedings of the 17th Australian Joint Conference on Advances in Artificial Intelligence*, pages 488–499.

Pintu Lohar, Koel Dutta Chowdhury, Haithem Afli, Mohammed Hasanuzzaman, and Andy Way. 2017. ADAPT at IJCNLP-2017 Task 4: A Multinomial Naive Bayes Classification Approach for Customer Feedback Analysis task. In *Proceedings of the IJCNLP 2017, Shared Tasks*, pages 161–169.

Andrew D McCallum and Kamal Nigam. 1998. A Comparison of Event Models for Naive Bayes Text Classification. In *Proceedings of ICML/AAAI 98 Workshop on Learning of Text Categorisation*, pages 41–48, Madison, Wisconsin.

Niloy Mukherjee, Braja Gopal Patra, Dipankar Das, and Sivaji Bandyopadhyay. 2016. JU-NLP at Semeval-2016 task 11: Identifying complex words in a sentence. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 986–990, San Diego, California.

Gustavo Paetzold and Lucia Specia. 2015. Lexenstein: A framework for lexical simplification. In *Proceedings of the System Demonstrations at the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing*, pages 85–90, Beijing, China.

Gustavo Paetzold and Lucia Specia. 2016. SemEval 2016 Task 11: Complex Word Identification. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 560–569, San Diego, California.

system	English Wiki	English News	English WikiNews	German Wiki	Spanish Wiki	French Wiki
best	81.2	87.9	84.3	76.2	78.4	80.2
nb24	**74.6 (16)**	82.8 (21)	75.4 (24)	70.9 (3)	72.6 (10)	/
nb24-allen	73.0 (20)	**83.5 (17)**	**77.2 (20)**	/	/	53.0 (8)
nb24-en	/	/	/	/	/	51.8 (9)
nb24-de	/	/	/	/	/	55.1 (7)
nb24-es	/	/	/	/	/	**59.8 (5)**
worst	34.1 (28)	17.6 (34)	56.8 (31)	58.1 (14)	701 (17)	51.8 (9)

Table 7: Official accurracies and ranks (in parenthesis) for English, German, Spanish and French test sets used in the CWI shared task 2018: the best system, all our submissions, and the worst system.

Maja Popović. 2015. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the 10th Workshop on Statistical Machine Translation (WMT 15)*, pages 392–395, Lisbon, Portugal.

Matthew Shardlow. 2013. A comparison of techniques to automatically identify complex words. In *Proceedings of the Student Research Workshop at the 51st Annual Meeting of the Association for Computational Linguistics (ACL 2013*, Sofia, Bulgaria.

Miloš Stanojević and Khalil Sima'an. 2014. BEER: BEtter Evaluation as Ranking. In *Proceedings of the Ninth Workshop on Statistical Machine Translation (WMT 14)*, pages 414–419, Baltimore, Maryland, USA.

Weiyue Wang, Jan-Thorsten Peter, Hendrik Rosendahl, and Hermann Ney. 2016. CharacTer: Translation Edit Rate on Character Level. In *Proceedings of the First Conference on Machine Translation (WMT 16)*, pages 505–510, Berlin, Germany.

Seid Muhie Yimam, Chris Biemann, Shervin Malmasi, Gustavo Paetzold, Lucia Specia, Sanja Štajner, Anaïs Tack, and Marcos Zampieri. 2018. A Report on the Complex Word Identification Shared Task 2018. In *Proceedings of the 13th Workshop on Innovative Use of NLP for Building Educational Applications*, New Orleans, United States.

Seid Muhie Yimam, Sanja Štajner, Martin Riedl, and Chris Biemann. 2017a. CWIG3G2 - Complex Word Identification Task across Three Text Genres and Two User Groups. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 401–407, Taipei, Taiwan.

Seid Muhie Yimam, Sanja Štajner, Martin Riedl, and Chris Biemann. 2017b. Multilingual and Cross-Lingual Complex Word Identification. In *Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2017*, pages 813–822, Varna, Bulgaria.

Marcos Zampieri, Shervin Malmasi, Gustavo Paetzold, and Lucia Specia. 2017. Complex Word Identification: Challenges in Data Annotation and System Performance. In *Proceedings of the 4th Workshop on Natural Language Processing Techniques for Educational Applications (NLPTEA 2017)*, pages 59–63, Taipei, Taiwan.

Marcos Zampieri, Liling Tan, and Josef van Genabith. 2016. MacSaar at SemEval-2016 Task 11: Zipfian and Character Features for Complex Word Identification. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 1001–1005, San Diego, California.