Research on the Evaluation of Token Imbalance Degree of NMT Corpus

Anonymous ACL submission

Abstract

As a kind of classifier, neural machine translation (NMT) is known to perform better with balanced tokens during training. Studying the token distribution in NMT corpus is of guiding significance to improve its quality and the translation effect. Due to the existing researches on token imbalance degree have deficiencies in algorithm performance and word segmentation scope, we propose the Dispersion of Token Distribution (DTD) algorithm, and use it to evaluate corpus from three segmentation levels: character, subword and word. Our experiments show that this algorithm has an improvement in accuracy, effectiveness and robustness. Meanwhile, we find that the token imbalance degree of NMT corpus varies greatly at different segmentation levels, among which character has the highest, word has the lowest and subword is in between. In addition, we also find the regularities of token imbalance degree in languages German (DE), English (EN), French (FR) and Russian (RU).

1 Introduction

As an important topic of the Natural Language Processing (NLP), NMT is developing rapidly. Since Cho et al. (2014) constructed the NMT model by using RNN Encoder-Decoder network, many researchers have proposed methods to improve its performance. For example, Sutskever et al. (2014) used the Sequence to Sequence method to improve the translation effect of long sentences. Bahdanau et al. (2014) increased the BLEU (Papineni et al., 2002) score by conducting joint learning to align and translate. Sennrich et al. (2016) effectively improved the translation effect of low-frequency words by using subword units. With the development of NMT model, the segmentation method of NMT corpus has changed a lot. Different from the traditional phrase-based statistical machine translation model (Koehn et al., 2003; Chiang, 2007), NMT model generally adopts word-level segmentation method which caters more to the characteristics of neural networks. However, it usually produces many low-frequency words and generates a large vocabulary size, which affects its translation performance. In order to alleviate this problem, some models using smaller token granularity have been proposed, such as the widely used Byte Pair Encoding (BPE) model (Sennrich et al., 2016), the hybrid word-character-based model (Luong and Manning, 2016) and the word-piece-based model (Wu et al., 2016). Character-level model (Lee et al., 2017; Cherry et al., 2018) can divide corpus into the smallest granularity, and greatly reduce the vocabulary size, which makes it have advantages in multilingual machine translation.

With different segmentation levels, the token distribution of NMT corpus varies greatly, but due to the Zipfian (Zipf, 1949) nature of language, the token imbalance phenomenon is inevitably existing. It will lead to the over-fitting of high frequency tokens and under-training of low frequency tokens, which affects the translation effect. Many researchers (Jiang et al., 2019; Gu et al., 2020) have tried to eliminate the adverse effects caused by this phenomenon. However, few have studied its extent in NMT corpus. Gowda and May (2020) adopted algorithms D and $F_{95\%}$ to evaluate the token imbalance degree in their study. However, their research has the following defects: (1) The score of algorithm D is between 0-1, and the results of different word segmentation levels and corpora vary slightly, which is not conducive to compare the token imbalance situation. In addition, we find it is not accurate and robust to measure the token imbalance degree. (2) Algorithm $F_{95\%}$ only counts the number of a special token in NMT corpus, which can not effectively evaluate the token imbalance degree. (3) They did not investigate the effect of different segmentation levels on the token imbalance degree of corpus. Aimed at these shortages, we do a lot of work, and the contributions of this paper are as follows:

(1) We propose the DTD algorithm to better calculate the token imbalance degree of NMT corpus. Compared with previous studies, this algorithm has an improvement in accuracy, validity and robustness.

(2) We extend the segmentation level from subword to the three most widely used levels in NMT: character, subword and word, and their token imbalance degree has the following rules: character > subword > word.

(3) By comparing the DTD values of different languages, we find the regularities of token imbalance degree in languages DE, EN, FR and RU.

2 Related Work

2.1 Related Background

The core part of NMT model is the Encoder-Decoder network whose structure is shown in Figure 1. Before training, source and target sentences

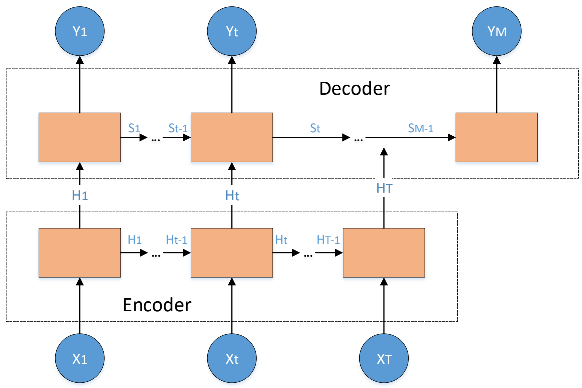


Figure 1: The structure of Encoder-Decoder network.

are divided into characters, subwords or words, then the generated tokens are converted into vectors by using word embedding technology (Mikolov et al., 2013). The sequence of source vectors is denoted as $X = (X_1, ..., X_T)$, the sequence of target vectors is denoted as $Y' = (Y'_1, ..., Y'_M)$, and the translation result is denoted as $Y = (Y_1, ..., Y_M)$. During training, the source vector $X_i (i \in [1, T])$ will be sent to the Encoder network one by one, and its output at time t is called the source hidden state H_t. The output of the Decoder network at time t is called the target hidden state S_t which can generate the translation vector Y_t through multi-layer neural network. The input of the Decoder network at time t is the source hidden state H_t and the target hidden state S_{t-1}. Judging by the output of the Decoder network, NMT is a multi-classifier, which is the reason why imbalanced tokens have negative effects on its performance. The optimization process of NMT model is completed by minimizing the loss of cross entropy, and the loss function is shown in Equation 1:

$$L = -\frac{1}{M}\sum_{j=1}^{M} \log P(Y|Y'_j < M, X) \quad (1)$$

As a parallel corpus, the two languages contained in NMT corpus are mutually source and target. As we can be seen from the above loss function, NMT model is optimized based on target tokens, so it is more reasonable to choose target corpus as the evaluation object of token imbalance degree in practical.

2.2 Related Algorithm

Token imbalance is a kind of class imbalance. Johnson and Khoshgoftaar (2019) systematically studied the class imbalance problem in deep learning, and introduced the algorithm ρ to represent the class imbalance level in their article. It is computed as:

$$\rho = \frac{\max_i\{|C_i|\}}{\min_i\{|C_i|\}} \quad (2)$$

In the above equation, C_i is a set of examples in class i, and $max_i\{|C_i|\}$ and $min_i\{|C_i|\}$ return the maximum and minimum class size over all i classes, respectively. When the class size is balanced, the ρ value is 1. The larger the ρ value is, the higher the class imbalance degree is.

Gowda and May (2020) used two algorithms to calculate the token imbalance degree of NMT corpus in their study. The first algorithm is called D, which is a simplified form of EMD distance (Rubner et al., 2000). It is used to count the sum of the frequency offsets of all the tokens and computed as:

$$D = \frac{1}{2}\sum_{i=1}^{k} |p_i - \frac{1}{K}|; \quad 0 \le D \le 1 \quad (3)$$

In Equation 3, K represents the number of token classes, and p_i represents the frequency of each token. When the frequencies of all the tokens in NMT corpus are equal, the D value is 0, which means the tokens are balanced. The larger the D value is, the higher the token imbalance degree is. The second algorithm is called $F_{95\%}$, and its principle is: First, all the tokens are sorted in order of number from high to low, then the number of the token ranked 95%-th (The author thinks that there are many impurities in the last 5% tokens, so they are not taken into account) is denoted as $F_{95\%}$. The larger the $F_{95\%}$ value is, the smaller the proportion of low frequency tokens in corpus is. We think algorithm $F_{95\%}$ has several defects: (1) It only calculates the number of the 95%-th token in NMT corpus, but does not consider the number difference between all the tokens. Therefore, it does not comprehensively reflect the token imbalance situation. (2) The algorithm has a parameter to set, the selection of 95% has no theoretical basis, and it is impossible to find a value that is suitable for all corpora. (3) Although the last 5% tokens contain some impurities, but also have some tokens that have important semantic information. And they are important indicator of the quality and token imbalance degree of NMT corpus and should not be excluded.

In order to evaluate the token imbalance degree of NMT corpus more comprehensively and accurately, we propose the DTD algorithm whose principle is as follows: Suppose there are n different tokens in a corpus, denoted as $X_i(i \in [1, n])$ and the number of each token X_i is expressed as C_i. For NMT model, ideally the training data is balanced, that is, $C_1 = C_2 = \ldots = C_n$. Therefore, we calculate the standard deviation of C_i as the evaluation criterion of token imbalance degree. Its complete calculation process is as follows:

1. Calculate the average value of C_i as $\bar{C}$.

$$\bar{C} = \frac{1}{n}\sum_{i=1}^{n} C_i \tag{4}$$

2. Calculate the standard deviation of C_i as the DTD value.

$$DTD = \sqrt{\frac{\sum_{i=1}^{n}(C_i - \bar{C})^2}{n}} \tag{5}$$

The number of tokens is counted from the entire corpus rather than a sample, so the denominator under the square root of the Equation 5 is n rather than n-1. When all the tokens in NMT corpus have the same number, the DTD value is 0. The larger the DTD value is, the higher the token imbalance degree is.

3 Experiments and Results

3.1 Data and Settings

We choose the **News-Commentary-v10** and **Common Crawl corpus** of **WMT15**[1] as our subjects. For each corpus, we select four languages: DE, EN(from DE-EN parallel corpus), FR and RU, and segment them from three levels: character, subword and word. Character-level segmentation directly divides corpus into the minimum granularity. Word-level uses space as the symbol to segment corpus. Subword-level segmentation uses the BPE algorithm to process corpus. BPE has a single hyperparameter named *merge operations* that governs the vocabulary size. If the merge operations is set too large, the segmentation effect is not obvious, while if the merge operations is set too small, part of the tokens will be discarded, which will affect the token imbalance evaluation. Therefore, we set the merge operations to be between the vocabulary size of character and word. The merge operations of News-Commentary-v10 and Common Crawl corpus are set to 30K and 200K, respectively.

3.2 Steps and Results

Before segmentation, we use the *normalize-punctuation*, *remove-non-printing-char* and *tokenizer* scripts of **Moses**[2] to preprocess the corpus. After that, we segment the corpus at character, subword and word levels, respectively, sort each generated token X_i according to its number C_i from low to high, and give it a serial number as X_i_id. Then, we plot the tokens of each corpus at three segmentation levels in the plane coordinate system with X_i_id as X-axis and C_i as Y-axis, as shown in appendix A. Here, we just show the token distribution of News-Commentary-v10 at character-level in Figure 2. In order to conveniently observe the differences of key information in the token distribution, we denote the vocabulary size as N, the maximum token number as Max[C_i], the number of tokens with size one as N', the ratio of N' to N as K, and summarize these data in Table 1. Finally, we calculate the ρ, D, $F_{95\%}$ and DTD values of each corpus at three segmentation levels, and show them in Table 2, 3, 4and 5, respectively.

[1] Available at http://www.statmt.org/wmt15/translation-task.html

[2] Available at https://github.com/moses-smt/mosesdecoder

Corpus	Language	Segmentation level	N	Max[C_i]	N'	K
News-Commentary-v10	DE	Character-level	268	5514163	38	14.18%
		Subword-level	29974	326955	292	0.97%
		Word-level	165231	327012	84401	51.08%
	EN	Character-level	297	5455143	36	12.12%
		Subword-level	29456	285476	688	2.34%
		Word-level	84573	285497	35219	41.64%
	FR	Character-level	206	6098699	20	9.71%
		Subword-level	29362	292188	624	2.13%
		Word-level	81960	291734	31863	38.88%
	RU	Character-level	240	4395297	18	7.50%
		Subword-level	30118	361576	239	0.79%
		Word-level	172275	361597	78863	45.78%
Common Crawl corpus	DE	Character-level	2850	54603260	982	34.42%
		Subword-level	202768	2853693	2470	1.22%
		Word-level	1786351	2853693	1077160	60.30%
	EN	Character-level	3140	58789669	1096	34.90%
		Subword-level	200291	2957144	4511	2.25%
		Word-level	953787	2956646	540866	56.71%
	FR	Character-level	2451	90154836	834	34.03%
		Subword-level	200306	4447357	3313	1.65%
		Word-level	1042401	4444928	562135	53.93%
	RU	Character-level	1915	20610711	562	29.30%
		Subword-level	199748	1367921	2773	1.39%
		Word-level	818213	1367921	436963	53.40%

Table 1: The N, Max[C_i], N' and K of each corpus at three segmentation levels. N represents the vocabulary size, Max[C_i] represents the maximum number of tokens, N' represents the number of tokens with size one, and K represents the ratio of N' to N.

Corpus	Language	Character-level	Subword-level	Word-level
News-Commentary-v10	DE	5514163	326955	327012
	EN	5455143	285476	285497
	FR	6098699	292188	291734
	RU	4395297	361576	361597
Common Crawl corpus	DE	54603260	2853693	2853693
	EN	58789669	2957144	2956646
	FR	90154836	4447357	4444928
	RU	20610711	1367921	1367921

Table 2: The ρ values of each corpus at three segmentation levels.

Corpus	Language	Character-level	Subword-level	Word-level
News-Commentary-v10	DE	0.837	0.661	0.835
	EN	0.864	0.724	0.837
	FR	0.835	0.740	0.841
	RU	0.824	0.601	0.790
Common Crawl corpus	DE	0.971	0.748	0.877
	EN	0.975	0.824	0.907
	FR	0.967	0.822	0.912
	RU	0.937	0.707	0.826

Table 3: The D values of each corpus at three segmentation levels.

Corpus	Language	Character-level	Subword-level	Word-level
News-Commentary-v10	DE	1	7	1
	EN	1	3	1
	FR	1	4	1
	RU	1	12	1
Common Crawl corpus	DE	1	9	1
	EN	1	4	1
	FR	1	5	1
	RU	1	6	1

Table 4: The $F_{95\%}$ values of each corpus at three segmentation levels.

Corpus	Language	Character-level	Subword-level	Word-level
News-Commentary-v10	DE	5.73e5	3.04e3	1.31e3
	EN	4.64e5	3.18e3	1.89e3
	FR	6.46e5	3.45e3	2.07e3
	RU	4.39e5	2.63e3	1.11e3
Common Crawl corpus	DE	1.67e6	1.05e4	3.64e3
	EN	1.53e6	1.27e4	5.94e3
	FR	2.76e6	1.92e4	8.49e3
	RU	6.67e5	4.14e3	2.12e3

Table 5: The DTD values of each corpus at three segmentation levels.

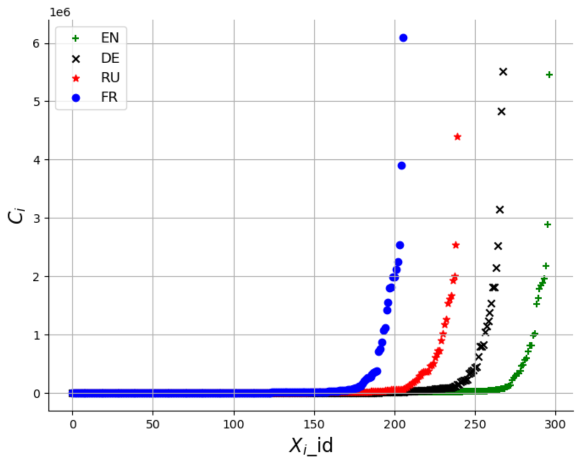


Figure 2: The token distribution of News-Commentary-V10 at character-level. The X-axis represents token order and the Y-axis represents token number.

4 Analysis

4.1 Algorithm Analysis

Algorithm ρ only calculates the ratio of the maximum token size to the minimum, without considering the size differences between all the tokens. Therefore, it may not correctly reflect the data imbalance degree. For example, suppose there are sets X=[1, 2], Y=[1, 1, 1, 1, 1, 2]. The ρ value of set X is equal to that of set Y, which means they have a same data imbalance degree, but we all know that the data of set Y is more balanced. By observing the N' values in Table 1, it can be seen that no matter which segmentation level is adopted, there are always some tokens with size one in the corpus. So the ρ values in Table 2 are only controlled by the Max[C_i] values in Table 1. In other words, the token imbalance degree is represented only by the maximum token number, which is not reasonable. Therefore, using algorithm ρ to evaluate the token imbalance degree of NMT corpus is not an advisable choice.

Algorithm D represents the sum of the frequency offsets of all the tokens in corpus, which to some extent reflects its token imbalance degree. As we can see from Table 3, due to the D value is between 0 and 1, except that the D value of subword is significantly smaller than that of character and word, the results vary slightly between different languages and segmentation levels, which is not conducive to compare the token imbalance situation. By carefully observing the data in Table 3, it can be found that the D values of most corpora at character-level are slightly larger than that at word-level, but the FR language of News-Commentary-v10 is a counter example. In addition, the token imbalance degrees of languages DE, EN and RU have the following regularity: EN > DE > RU, while language FR does not. The above facts indicate that the regularity of algorithm D in terms of segmentation level and language is easily influenced by the corpus, which means it is not robust enough.

Algorithm $F_{95\%}$ only counts the token number in NMT corpus, which can hardly reflect the token imbalance degree. In Table 4, we can find that the $F_{95\%}$ values of all corpora at character and word levels are 1, which makes it impossible to judge the token imbalance situation and to compare the degree between different languages and segmentation levels. The $F_{95\%}$ values of corpora at subword-level are higher than that at character and word levels, indicating that subword-level segmentation can reduce the proportion of low-frequency tokens in corpus, but it does not mean that the token imbalance degree of subword is definitely higher or lower than that of character and word.

Algorithm DTD represents the dispersion of token number in NMT corpus, which accurately reflects the token imbalance degree. Table 5 shows that the DTD values vary significantly between different word segmentation levels and languages with strong regularity. For example, the DTD value of character is about two orders of magnitude higher than that of subword which is about one times larger than word. In addition, when using subwords and words, the token imbalance order of the four languages is FR > EN > DE > RU, and when using characters, the order is FR > DE > EN > RU. Therefore, compared with algorithm D, DTD algorithm shows stronger regularity and better robustness in terms of segmentation level and language.

The data in Table 5 show that the DTD values of subword are larger than that of word, indicating that it leads to a higher token imbalance degree. The data in Table 3 show that the D values of subword are smaller than that of word, which indicates that it alleviates the token imbalance phenomenon. The conclusions of these two algorithms are contradictory. To figure out which algorithm is right, we conduct a further analysis. Suppose there is a corpus A, whose word distribution is 1 "desk", 2 "taller", 2 "cheaper", 7 "tall", 7 "cheap" and 10 "stronger". Since corpus A contains a large number of "er", if we segment it at subword-level and set the vocabulary size to 5, the token distribution will be 1 "desk", 9 "tall", 9 "cheap", 10 "strong" and 14 "er". The DTD value of word is: DTD1 = DTD[1,2,2,7,7,10] = 3.34, and that of subword is: DTD2 = DTD[1,9,9,10,14] = 4.22. The results indicate that subword-level segmentation does increase the DTD value compared with word. The D value of word is: D1 = D[1,2,2,7,7,10] = 0.328, and that of subword is: D2 = D[1,9,9,10,14] = 0.177. The results show that subword-level segmentation indeed reduce the D value compared with word. The conclusion of corpus A is consistent with that of algorithms DTD and D. Then, we sort its subwords and words in order of number from low to high, and draw them in the plane coordinate system, as shown in Figure 3. When the tokens of a corpus have the same number, its token distribution in plane coordinate system is a horizontal line. The more slant the distribution line is, the higher the token imbalance degree is. In Figure 3, it can be

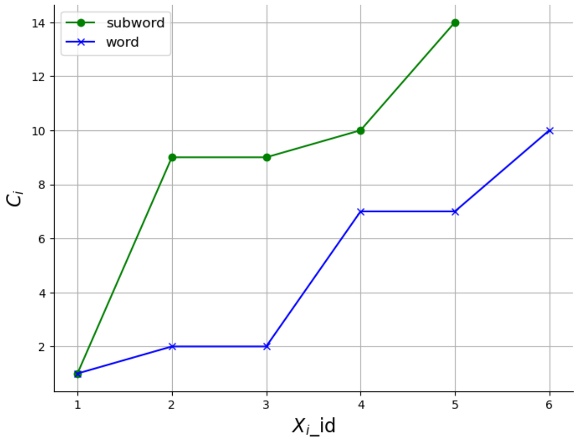


Figure 3: The token distribution of corpus A at subword and word levels. The X-axis represents token order and the Y-axis represents token number.

easily seen that the overall trend of distribution line of subword is more inclined than that of word, indicating that subword-level segmentation can lead to a more imbalanced tokens. Therefore, the conclusion of algorithm D is wrong, which verifies that it is not as accurate as algorithm DTD in measuring the token imbalance degree of NMT corpus.

Through the above analysis of the four algorithms, it can be seen that the DTD algorithm has better accuracy and robustness and can reflect the token imbalance degree more comprehensively and effectively.

4.2 Word Segmentation Level Analysis

Studying the token imbalance degree of NMT corpus at different segmentation levels is helpful to the selection of appropriate segmentation method for NMT model. For this purpose, we conduct the following analysis.

In Table 5, it can be seen that the token imbalance degree of NMT corpus at different segmentation levels has the following pattern: character > subword > word, which indicates that the token imbalance degree of character is higher than that of subword which is higher than word. Some researchers (Gowda and May, 2020; Gu et al., 2020) believe that compared with word, the use of subwords can improve the translation effect of low-frequency tokens, indicating that the token imbalance situation is alleviated, but we show that the opposite is true. However, it is hard to explain it only by observing the token distribution, so we try to figure it out based on the key data in Table 1.

By observing the N and Max[C_i] values in Table 1, it can be seen that the vocabulary size of character decreases significantly compared with subword, and the maximum token number increases greatly. In addition, by observing the N' values in Table 1, it can be found that compared with subword, there are still some tokens with size one in character, and the number decreases. The token distribution line of corpus is close to the exponential function, which Zipf (1949) had already found. Therefore, compared with subword, the overall trend of distribution line of character is steeper (The transverse span of exponential function becomes smaller, the maximum value gets bigger, and the minimum value remains unchanged), thus its token imbalance degree is higher.

The experimental results of corpus A have shown that subword-level segmentation can increase the inclination of token distribution line compared with word. We verify it again by analyzing the data in Table 1. Compared with word, the vocabulary size N of subword is significantly reduced, the maximum token number Max[C_i] is almost unchanged (The FR language of News-Commentary-v10 changes the most, only increasing by 0.156%.), and there are still some tokens with size one (The proportion K and number N' are both significantly reduced.). Therefore, the overall trend of its token distribution line is steeper (The transverse span of exponential function becomes smaller, the maximum value is almost unchanged, and the minimum value remains unchanged.), which means a more imbalanced token. If token "desk" is not contained in corpus A, the DTD value of word is: DTD1' = DTD[2,2,7,7,10] = 3.14, and that of subword is: DTD2' = DTD[9,9,10,14] = 2.06. The results show that subword-level segmentation can reduce the token imbalance degree in special cases where there is no token with size one. Therefore, it is the low frequency tokens that cannot be decomposed that increase the token imbalance degree of subword.

Compared with word-level segmentation, subword increases the token imbalance degree of NMT corpus, which will reduce the translation effect of some low-frequency tokens. For example, for corpus A, although using subwords improves the translation effect of "taller" and "cheaper", but it causes the token "desk" to be more under-training and have worse translation effect. However, subword-level segmentation also has its advantages. As shown by the K values in Table 1, with word-level segmentation, there are about 40%-60% tokens with size one, which means that there are a large proportion of low-frequency tokens in NMT corpus. Subword-level can greatly reduce this proportion to about 0.8% to 2.4%. Through the use of subword units, the vocabulary size and the proportion of low-frequency tokens are both effectively reduced. Even though the translation effect of some low-frequency tokens will be affected, the overall performance of NMT model will be improved.

4.3 Corpus and Language Analysis

Studying the token imbalance degree between different corpora and languages is conducive to select a higher quality corpus and improve the multilingual translation performance, which is the significance of this subsection. By observing the data in Table 5, we can find that for the same language, the DTD value of Common Crawl corpus is larger than that of News-Commentary-v10, indicating that the News-Commentary-v10 corpus has a smaller token imbalance degree. This conclusion is consistent with our expectations, because the content of News-Commentary-v10 only involves the field of news commentary, which makes it doesn't have too many token classes. In addition, news commentary has strict requirements for language accuracy, so the corresponding corpus is of high quality. By contrast, the content of Common Crawl corpus is directly crawled from the web, and involves many fields, which leads to a large number of token classes. And its quality cannot be strictly controlled, resulting in a lot of impurities in the corpus, which affects its token imbalance degree. From the perspective of language, when using subword-level and word-level segmentation, the order of token imbalance degree is FR > EN > DE > RU. When character-level segmentation is used, the order becomes FR > DE > EN > RU. We think these patterns are related to the unique Zipfian (Zipf, 1949) nature of each language. Due to the limited space, we don't carry out further research which will be our future work.

5 Conclusion

There are many researches aimed at solving the adverse effects of token imbalance, but few have evaluated its degree in NMT corpus, and there are some shortages in these existing studies. Aimed at these shortages, this paper proposes the DTD algorithm and uses it to analyze different corpora from character, subword and word segmentation levels. Experimental results show that the proposed algorithm has better accuracy, effectiveness and robustness than previous studies. By comparing the DTD value of NMT corpus at different segmentation levels, this paper finds that character has the highest token imbalance degree, word has the lowest and subword is in between, and the view of using subwords can alleviate the token imbalance degree compared with word is proved wrong in this paper. In addition, by comparing the results of different languages, this paper also finds that languages DE, EN, FR and RU have regularity in token imbalance degree, which could be related to the characteristics of each language.

In future work, we will apply the algorithm proposed in this paper to more corpora and languages to obtain more valuable findings. In addition, we will also focus on studying and solving the adverse effects of the token imbalance problem in NMT corpus.

References

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural Machine Translation by Jointly Learning to Align and Translate. *arXiv e-prints*, page arXiv:1409.0473.

Colin Cherry, George Foster, Ankur Bapna, Orhan Firat, and Wolfgang Macherey. 2018. Revisiting character-based neural machine translation with capacity and compression. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4295–4305, Brussels, Belgium. Association for Computational Linguistics.

David Chiang. 2007. Hierarchical Phrase-Based Translation. *Computational Linguistics*, 33(2):201–228.

Kyunghyun Cho, Bart van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using RNN encoder–decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1724–1734, Doha, Qatar. Association for Computational Linguistics.

Thamme Gowda and Jonathan May. 2020. Finding the optimal vocabulary size for neural machine translation. *Findings of the Association for Computational Linguistics: EMNLP 2020*.

Shuhao Gu, Jinchao Zhang, Fandong Meng, Yang Feng, Wanying Xie, Jie Zhou, and Dong Yu. 2020. Token-level adaptive training for neural machine translation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1035–1046, Online. Association for Computational Linguistics.

Shaojie Jiang, Pengjie Ren, Christof Monz, and Maarten de Rijke. 2019. Improving neural response diversity with frequency-aware cross-entropy loss. *The World Wide Web Conference on - WWW '19*.

Justin Johnson and Taghi Khoshgoftaar. 2019. Survey on deep learning with class imbalance. *Journal of Big Data*, 6:27.

Philipp Koehn, Franz J. Och, and Daniel Marcu. 2003. Statistical phrase-based translation. In *Proceedings of the 2003 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*, pages 127–133.

Jason Lee, Kyunghyun Cho, and Thomas Hofmann. 2017. Fully character-level neural machine translation without explicit segmentation. *Transactions of the Association for Computational Linguistics*, 5:365–378.

Minh-Thang Luong and Christopher D. Manning. 2016. Achieving open vocabulary neural machine translation with hybrid word-character models. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1054–1063, Berlin, Germany. Association for Computational Linguistics.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, page 311–318, USA. Association for Computational Linguistics.

Yossi Rubner, Carlo Tomasi, and Leonidas J. Guibas. 2000. The earth mover's distance as a metric for image retrieval. *Int. J. Comput. Vision*, 40(2):99–121.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany. Association for Computational Linguistics.

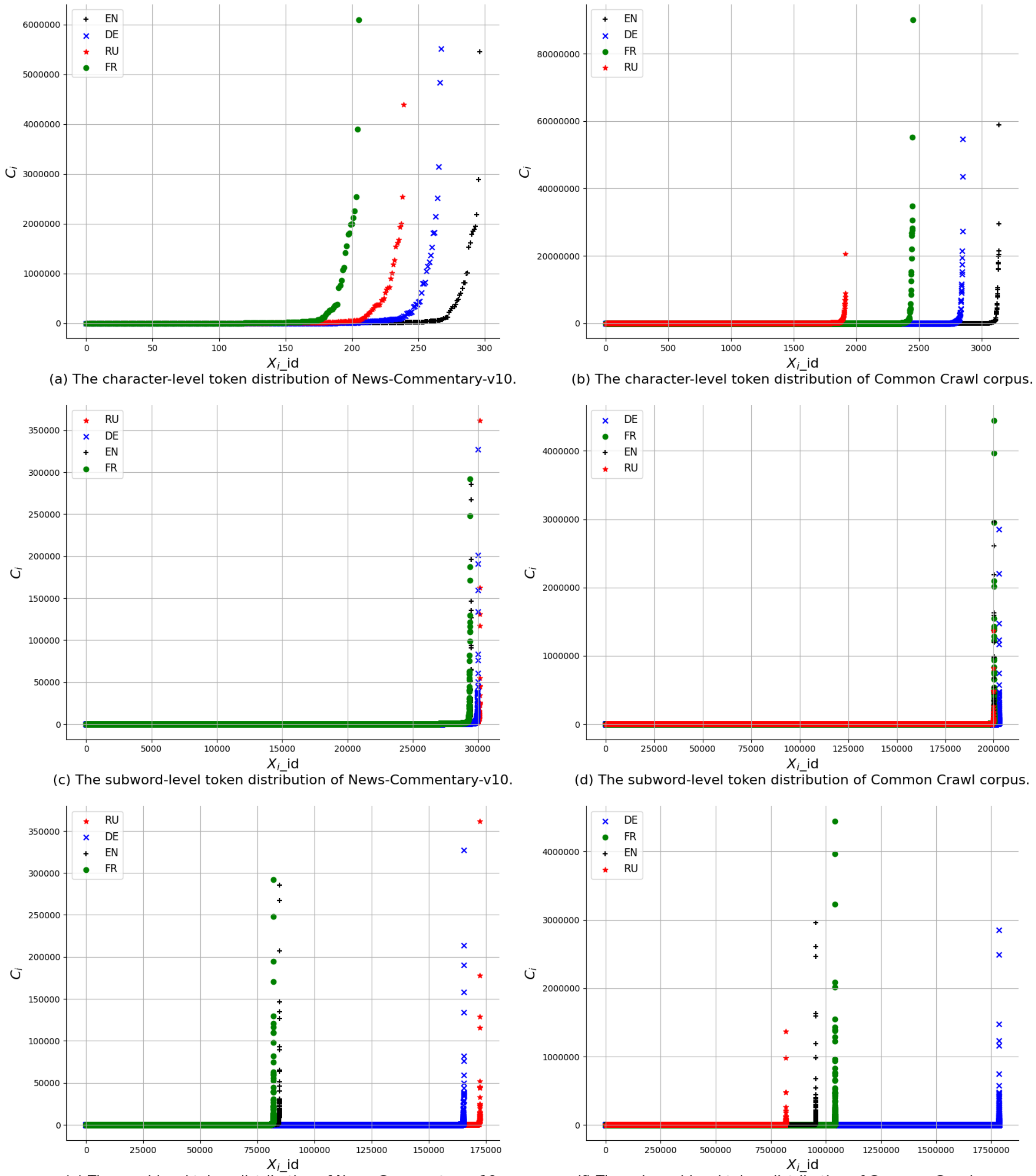
Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. 2014. Sequence to sequence learning with neural networks.

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V. Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, Jeff Klingner, Apurva Shah, Melvin Johnson, Xiaobing Liu, Łukasz Kaiser, Stephan Gouws, Yoshikiyo Kato, Taku Kudo, Hideto Kazawa, Keith Stevens, George Kurian, Nishant Patil, Wei Wang, Cliff Young, Jason Smith, Jason Riesa, Alex Rudnick, Oriol Vinyals, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2016. Google's neural machine translation system: Bridging the gap between human and machine translation.

G. Zipf. 1949. Human behavior and the principle of least effort. 6.

A Token Distribution Figure

In Figure 4 we plot the token distribution of News-Commentary-V10 and Common Crawl corpus at there segmentation levels. In each subfigure, languages DE, EN, FR and RU are marked discriminatively.



(a) The character-level token distribution of News-Commentary-v10.

(b) The character-level token distribution of Common Crawl corpus.

(c) The subword-level token distribution of News-Commentary-v10.

(d) The subword-level token distribution of Common Crawl corpus.

(e) The word-level token distribution of News-Commentary-v10.

(f) The subword-level token distribution of Common Crawl corpus.

Figure 4: The token distribution of News-Commentary-V10 and Common Crawl corpus at there segmentation levels. The X-axis represents token order and the Y-axis represents token number.