NMT语料库中语符不平衡度的测评研究

王海波[1],余丽丽[2],王宏伟[3]
(1. 浙江大学生物医学工程与仪器科学学院,浙江杭州 310027;2. 浙江师范大学教师教育学院,浙江金华 321004;
3. 浙江大学伊利诺伊大学厄巴纳香槟校区联合学院,浙江海宁 314499)

摘 要: 语符不平衡是神经机器翻译(Neural Machine Translation,NMT)语料库中普遍存在的现象.评估NMT语料库的语符不平衡度对提升语料库质量和翻译效果具有重要意义.针对现有的语符不平衡度测评研究在算法和分词范围上的缺陷与不足,本文提出语符分布离散度算法(Dispersion of Token Distribution,DTD),用以计算语符不平衡度,并扩大分词范围,从字符、子词和词3种粒度对语料库进行评估.实验结果表明,该算法在准确度、有效性和鲁棒性方面较以往研究有较大提升;语料库在不同分词粒度下的语符不平衡度差异很大,其中字符粒度的语符不平衡度最大,子词粒度次之,词粒度最小.

关键词: 神经机器翻译;语料库;分词;粒度;语符不平衡度
基金项目: 国家重点研发计划(No.2020YFB1707803);浙江大学科研资助项目(No.XY2021018)
中图分类号: TP391 **文献标识码:** A **文章编号:** 0372-2112(2023)10-2884-10
电子学报URL:http://www.ejournal.org.cn **DOI:**10.12263/DZXB.20211369

Research on Evaluation of Token Imbalance Degree in NMT Corpus

WANG Hai-bo[1], YU Li-li[2], WANG Hong-wei[3]
(1. *College of Biomedical Engineering & Instrument Science, Zhejiang University, Hangzhou, Zhejiang* 310027, *China*;
2. *College of Teacher Education, Zhejiang Normal University, Jinhua, Zhejiang* 321004, *China*;
3. *ZJU-UIUC Joint Institute, Zhejiang University, Haining, Zhejiang* 314499, *China*)

Abstract: Token imbalance is a common phenomenon in the corpus of neural machine translation (NMT). It is of great significance to evaluate the token imbalance degree of NMT corpus to improve the quality of corpus and translation effect. Aiming at the defects and deficiencies in the algorithm and word segmentation scope of the existing studies on the measurement of the token imbalance degree, this paper proposes the dispersion of token distribution (DTD) algorithm to calculate the token imbalance degree, expands the word segmentation scope, and evaluates the corpus from three granularity: character, subword and word. The experimental results show that the accuracy, validity and robustness of the proposed algorithm are greatly improved compared with previous studies. There are great differences in the token imbalance degree of corpora under different word segmentation granularity, in which character granularity has the highest token imbalance degree, followed by subword granularity and word granularity.

Key words: neural machine translation; corpus; word segmentation; granularity; token imbalance degree
Foundation Item(s): National Key Research and Development of China (No.2020YFB1707803); Zhejiang University Research Project (No.XY2021018)

1 引言

作为自然语言处理(Natural Language Processing,NLP)领域中的一个重要课题,NMT的发展十分迅速.自Cho等[1]在2014年首次使用RNN编码-解码网络构建NMT模型以来,不断有研究人员对其进行改进和优化.例如,Sutskever等人[2]使用“Sequence to Sequence”方法提升了长语句的翻译效果;Bahdanau等人[3]通过对齐和翻译的联合学习方式大大提高了模型的Bleu[4]分数;Sennrich等人[5]利用子词单元有效改善了低频词的翻译效果.此外,NMT语料库的分词方法也发生了很大变化.与传统的基于短语的统计机器翻译模型[6,7]不

收稿日期:2021-10-09;修回日期:2022-05-10;责任编辑:王天慧

同,NMT模型一般采用词粒度这种更加符合神经网络特点的分词方式. 然而,这种粒度的分词通常会产生大量的低频词,导致生成的词表规模极大,使NMT模型在训练和应用时内存消耗很大,进而影响模型的翻译性能. 为了有效减小词表规模,提高低频词的翻译效果,一些采用更小粒度分词方法的模型被提出,如广泛使用的字节对编码(Byte Pair Encoding,BPE)模型[5]、基于字和字符的混合模型[8]以及基于词碎片的模型[9]. 基于字符的模型[10,11]可以将语料分割为最小的粒度,极大地减小词表规模,使其在多语种机器翻译中独具优势,被越来越多的相关研究所采用.

无论使用哪种粒度的分词方法,NMT语料库中都会不可避免地出现语符不平衡现象. 这一现象会造成NMT模型在训练过程中,高频语符过拟合,低频语符欠拟合,进而影响其翻译效果. 已有的大多研究都致力于解决语符不平衡产生的不利影响,而涉及语料库自身的语符不平衡度测评的研究却很少. 因此,为提升语料库的质量和NMT翻译效果,深入研究语料库的语符不平衡度有着重要的现实意义. Gowda等人[12]系统地研究了NMT语料库的语符不平衡度,但是存在两点不足:(1)提出的语符不平衡度算法缺乏准确度、有效性和鲁棒性;(2)仅研究语料库在子词粒度下的语符不平衡度,而常用的分词粒度有字符、子词和词3种. 现针对这些缺陷和不足,本文提出一种更优的语符不平衡度算法——DTD,并将分词范围从子词粒度扩展到字符、子词和词3种粒度来对NMT语料库的语符不平衡度进行评估,得到了一些重要结论.

2　相关工作

2.1　相关背景

作为NMT模型的核心,编码-解码(Encoder-Decoder)网络的结构如图1所示,其工作原理如下. 在训练开始前,源语句和目标语句被分割为字符级、子词级或词级的语符,分词后产生的语符再通过词嵌入技术[13,14]转换为向量. 转换后的源语句的向量序列为 $\boldsymbol{X}=(\boldsymbol{X}_1,\boldsymbol{X}_2,\cdots,\boldsymbol{X}_T)$,其中 $\boldsymbol{X}_i$ ($i\in[1,T]$)为源语句的第 i 个语符向量, T 为源语句中语符的个数;目标语句的向量序列为 $\boldsymbol{Y}'=(\boldsymbol{Y}'_1,\boldsymbol{Y}'_2,\cdots,\boldsymbol{Y}'_M)$,其中 $\boldsymbol{Y}'_j$ ($j\in[1,M]$)为目标语句的第 j 个语符向量, M 为目标语句中语符的个数;翻译结果的向量序列为 $\boldsymbol{Y}=(\boldsymbol{Y}_1,\boldsymbol{Y}_2,\cdots,\boldsymbol{Y}_M)$,其中 $\boldsymbol{Y}_j$ ($j\in[1,M]$)为翻译结果的第 j 个语符向量, M 为翻译结果中语符的个数. 训练开始时,源语符向量 $\boldsymbol{X}_i$ 被逐项输入编码网络中进行编码,编码网络 t 时刻的输出称为源隐藏状态 $\boldsymbol{H}_t$,解码网络 t 时刻的输出称为目标隐藏状态 $\boldsymbol{S}_t$. 经多层神经网络生成翻译语符向量 $\boldsymbol{Y}_t$,解码网络 t 时刻的输入是当前时刻的源隐藏状态 $\boldsymbol{H}_t$ 和上一时刻的目标隐藏状态 $\boldsymbol{S}_{t-1}$. 从解码网络的输出形式来看,NMT本质上是一种多分类模型,这也是语符类别不平衡对其翻译效果具有不利影响的原因. NMT模型训练的优化过程是通过最小化交叉熵损失来完成的,损失函数为

$$L=-\frac{1}{M}\sum_{j=1}^{M}\log P\left(Y'|Y_j<M,X\right) \tag{1}$$

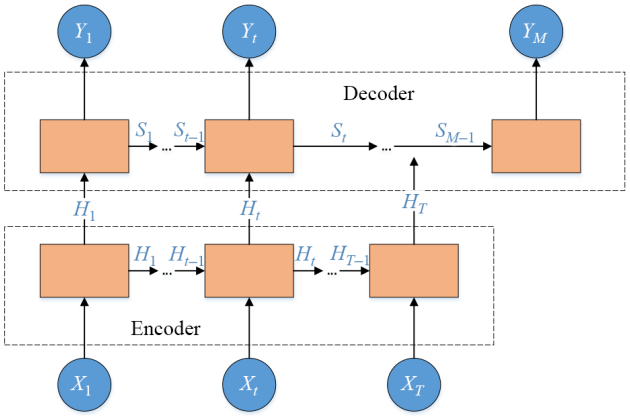


图1　NMT编码-解码网络结构

NMT语料库作为平行语料库,其包含的两种语言互为源语料库和目标语料库. 从上述损失函数可以看出,NMT模型是基于目标语料库的语符来进行优化的,所以在实际应用时选择目标语料库作为语符不平衡度的评估对象更为合理.

2.2　相关研究

NMT作为一种多分类模型[15,16],其语料库的语符不平衡问题属于数据的类别不平衡问题. 针对分类模型的数据不平衡问题,许多学者提出了解决办法[17~19]. 陆克中等人[20]系统地研究了深度学习中的数据类别不平衡现象,并在其文章中引入算法 ρ 来计算数据的类别不平衡度,其公式如下:

$$\rho=\frac{\max\left\{\left|C_i\right|\right\}}{\min\left\{\left|C_i\right|\right\}} \tag{2}$$

式(2)中, C_i 表示数据中类别为 i 的数据集合, $\max\{|C_i|\}$ 和 $\min\{|C_i|\}$ 分别表示在所有类别中的最大类别规模和最小类别规模. 当所有的类别具有相同的数据规模时,数据类别是平衡的,ρ 值为1. ρ 值越大,表明数据的类别不平衡度越大.

Gowda等人[12]在他们的研究中使用两种算法来表示NMT语料库的语符不平衡度. 第一种算法称为 D,它是EMD距离[21]的一种简化形式,用以统计所有语符的概率偏移之和,其式如下:

$$D=\frac{1}{2}\sum_{i=1}^{K}\left|p_i-\frac{1}{K}\right|;\ 0\leqslant D\leqslant 1 \tag{3}$$

式(3)中,K表示语料库中语符的类别总数,p_i表示每一种语符在语料库中出现的概率.因为所有语符的概率量之和为1,任何一种语符的概率量移出都会移入到其他语符的概率量中,所以为了避免重复统计,作者将概率偏移之和的计算结果除以2.当语料库中所有语符出现的概率相同时,语符分布是均衡的,D值为0.D值越大,表示语符不平衡度越大.算法D的计算结果为0~1,语料库在不同分词粒度下的结果差异细微,不利于进行对比分析.此外,通过实验还发现,算法D用以衡量语符不平衡度时缺乏准确度和鲁棒性.

第二种算法称为$F_{95\%}$,其原理是如下.首先将所有语符按照频数从高到低进行排序,然后统计频数排在第95%位语符的频数大小记为$F_{95\%}$.作者认为末尾的5%语符中包含许多杂质,因而不考虑在内.$F_{95\%}$数值越大,说明语料库的低频语符所占比例越小.算法$F_{95\%}$的计算量非常小,但其原理存在很多缺陷.首先,它只计算了语料库中第95%位语符的频数,而没有考虑高低频语符之间频数的差异大小,严格意义上说,并没有真正计算语料库的语符不平衡度;其次,95%这一参数的选取是随机的,没有理论依据,也并不适用于所有的语料库;最后,尽管末尾的5%语符中含有一些杂质,但也包含一些具有重要语义信息的语符.而且,这末尾的5%语符是体现语料库质量的一个重要方面,也是反映语料库语符不平衡度的一个重要指标,应该纳入计算范围之内.

本文提出的DTD算法原理如下.假设一个语料库中有n个不同的语符,记为$X_i(i\in[1,n])$,每个语符X_i的频数表示为$C_i(i\in[1,n])$.对于NMT模型来说,理想情况下每个语符的频数是相等的,即$C_1=C_2=\cdots=C_n$.然而,由于自然语言具有Zipfian[22]特性,语符的频数分布是离散的,即存在语符不平衡现象,因此,DTD算法通过计算语符频数C_i分布的标准差作为语符不平衡度的评估标准.具体算法如下:

(1)计算语符频数的平均值$\bar{C}$,即

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

(2)计算语符分布离散度DTD,即

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

语符频数C_i是从整个语料库中而不是样本中统计的,因此,式(5)平方根下的分母是n而不是$n-1$.当语料库中的所有语符具有相同频数时,语符分布处于平衡状态,DTD值为0.DTD值越大,表明语符不平衡度越大.利用DTD算法,可以准确量化NMT语料库中的语符不平衡度.语料库在不同分词粒度下的语符分布差异很大.因此,本文从字符、子词和词3种最常用的分词粒度来评估语料库的语符不平衡情况.

3 实验与结果

3.1 实验数据与参数设置

本文选择国际机器翻译大赛WMT15的News-Commentary-v10和Common Crawl corpus以及WMT21的News-Commentary-v16语料库作为数据集.对于每个语料库,选择DE,FR,EN(选自DE-EN平行语料库)和RU 4种语言作为评估对象,并从字符、子词和词3种粒度来对语料进行分词.其中,字符级分词直接按照最小粒度对语料库进行分割;词级分词以空格为分隔符对语料库进行分割;子词级分词使用BPE算法对语料库进行分割.对于BPE算法,词表规模选择过大会导致分词效果不明显,词表规模选择过小则会导致部分语符丢失,影响语符不平衡度评估结果.因此,实验将子词粒度的词表规模设置在介于字符粒度和词粒度之间.其中,News-Commentary-v10和News-Commentary-v16语料库的词表规模设置为30K,Common Crawl corpus语料库的词表规模设置为200K.

3.2 实验步骤与结果

在对语料库进行分词之前,首先使用Moses的normalize-punctuation,remove-non-printing-char以及tokenizer对语料库进行预处理,目的是将标点符号规范化,删除无法打印的字符,以及标记语料库中的所有语符.数据预处理后,对语料库分别进行字符级、子词级和词级的分词,将分词后产生的语符X_i按照其频数C_i从小到大进行排序,并赋予其序号X_i_id.然后,以序号X_i_id为X轴,频数C_i为Y轴,将3个语料库在3种分词粒度下的语符分布绘制在平面坐标系中,如图2所示.为了更直观地发现不同语料库在不同分词粒度下的语符分布在关键参数上的差异,本文将词表大小表示为N,语符频数的最大值表示为$\mathrm{Max}[C_i]$,频数为1的语符类别个数表示为N^*,N^*与N的比值表示为K,并将这些结果统计在表1中.最后,分别计算语料库在字符、子词和词3种分词粒度下的ρ,D,$F_{95\%}$以及DTD值,结果分别如表2~5所示.

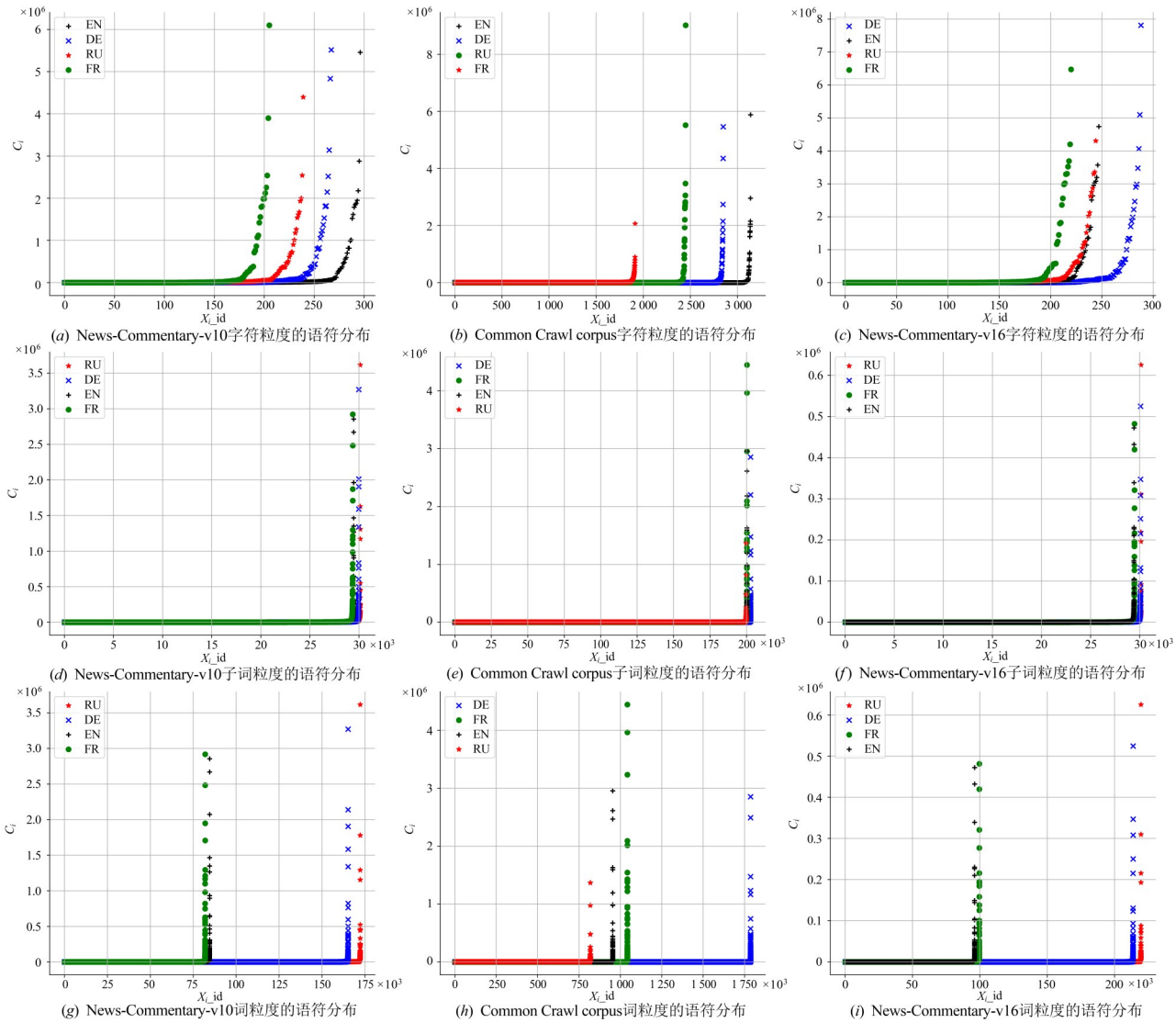


图2　3个语料库在3种分词粒度上的语符分布

4　实验结果分析

4.1　算法对比分析

ρ算法表示语料库中的最大语符频数与最小语符频数的比值. 通过观察表1的N^*值可知,不管是哪个语料库采用何种分词方法,语料库中总是含有频数为1的语符. 所以表2中采用ρ算法计算的语符不平衡度实际为语料库的最大语符频数,这显然无法对语料库进行有效的评估.

D算法表示语料库所有语符的概率偏移总和,一定程度上反映了语料库的语符不平衡度. 通过观察表3可以发现,由于D值为0~1,除了子词粒度的D值明显小于字符粒度和词粒度外,同一语料库的不同语言之间以及字符粒度和词粒度之间的D值差异细微,不利于直观比较. 再深入分析表3中数据可以发现,语料库在3种分词粒度间的D值不具有规律性. 此外,DE,EN,RU 3种语言在语符不平衡度上具有EN>DE>RU的规律,而FR语言则没有. 上述结果表明,算法D的实验结果在分词粒度和语言上的规律性受语料库的影响较大,其鲁棒性欠佳.

$F_{95\%}$算法仅统计语料库的低频语符的频数大小,无法有效反映语料库的语符不平衡度. 通过观察表4可以发现,所有语料库在字符粒度和词粒度上的$F_{95\%}$值均为1,无法判断语符不平衡度大小,更无法进行语符不平衡度的比较. 表4数据显示子词粒度的$F_{95\%}$值比字符粒度和词粒度略高,说明子词级分词可以减小语料库中低频语符的比例,但并不能说明子词级分词的语符不平衡度就一定高于或低于字符级和词级,还需要进一步地计算分析.

DTD算法表示语料库中所有语符的频数离散度,客观准确地反映了语符的不平衡度. 通过观察表5可以发现,不同分词粒度间和不同语言间的DTD值差异

表1 3个语料库在3种分词粒度上的N,Max[C_i],N^*和K值

语料库	语言	分词粒度	N	Max[C_i]	N^*	K
News-Commentary-v10	DE	字符级	268	5 514 163	38	14.18%
		子词级	29 974	326 955	292	0.97%
		词级	165 231	327 012	84 401	51.08%
	EN	字符级	297	5 455 143	36	12.12%
		子词级	29 456	285 476	688	2.34%
		词级	84 573	285 497	35 219	41.64%
	FR	字符级	206	6 098 699	20	9.71%
		子词级	29 362	292 188	624	2.13%
		词级	81 960	291 734	31 863	38.88%
	RU	字符级	240	4 395 297	18	7.50%
		子词级	30 118	361 576	239	0.79%
		词级	172 275	361 597	78 863	45.78%
Common Crawl corpus	DE	字符级	2 850	54 603 260	982	34.42%
		子词级	202 768	2 853 693	2 470	1.22%
		词级	1 786 351	2 853 693	1 077 160	60.30%
	EN	字符级	3 140	58 789 669	1 096	34.90%
		子词级	200 291	2 957 144	4 511	2.25%
		词级	953 787	2 956 646	540 866	56.71%
	FR	字符级	2 451	90 154 836	834	34.03%
		子词级	200 306	4 447 357	3 313	1.65%
		词级	1 042 401	4 444 928	562 135	53.93%
	RU	字符级	1 915	20 610 711	562	29.30%
		子词级	199 748	1 367 921	2 773	1.39%
		词级	818 213	1 367 921	436 963	53.40%
News-Commentary-v16	DE	字符级	289	7 813 350	52	17.99%
		子词级	30 084	524 502	255	0.85%
		词级	213 787	524 502	109 660	51.29%
	EN	字符级	248	4 736 665	44	17.74%
		子词级	29 421	472 025	607	2.06%
		词级	96 107	472 025	39 550	41.15%
	FR	字符级	221	6 470 617	24	10.86%
		子词级	29 471	481 821	508	1.72%
		词级	99 579	481 156	38 381	38.54%
	RU	字符级	245	4 302 888	23	9.39%
		子词级	30 124	625 074	200	0.66%
		词级	219 604	625 074	97 393	44.35%

显著,且具有以下规律:字符粒度>子词粒度>词粒度.此外,当使用子词级和词级分词时,4种语言的语符不平衡度顺序为FR > EN > DE > RU. 当使用字符级分词时,4种语言的语符不平衡度顺序为FR > DE > EN > RU. 这说明,相比于D算法,DTD算法在分词粒度和语言上展现的规律性更强,鲁棒性更好.

表5中的数据显示子词粒度的DTD值大于词粒度,说明词粒度的语符不平衡度更小;而表3中的数据显示,子词粒度的D值小于词粒度,又说明子词粒度的语符不平衡度更小,两种算法得出的结论是相互矛盾的. 为此,假设存在一个语料库A,其词粒度语符分布为1个“desk”,2个“taller”,3个“cheaper”,4个“tall”,5个“cheap”. 由于语料库A中含有多个“er”语符,如果对此语料库进行子词级分词(词表规模设置为4),则语符分布为1个‘desk’,5个‘er’,6个‘tall’,8个‘cheap’. 语料库A在词粒度上的DTD值为$\mathrm{DTD}_1 = \mathrm{DTD}[1,2,3,4,5] = 1.41$,在子词粒度上的DTD值为$\mathrm{DTD}_2 = \mathrm{DTD}[1,5,6,8] = 2.55$. 这表明子词级分词相比词级分词会增

表2　3个语料库在3种分词粒度上的ρ值

语料库	语言	字符粒度	子词粒度	词粒度
News-Commentary-v10	DE	5 514 163	326 955	327 012
	EN	5 455 143	285 476	285 497
	FR	6 098 699	292 188	291 734
	RU	4 395 297	361 576	361 597
Common Crawl corpus	DE	54 603 260	2 853 693	2 853 693
	EN	58 789 669	2 957 144	2 956 646
	FR	90 154 836	4 447 357	4 444 928
	RU	20 610 711	1 367 921	1 367 921
News-Commentary-v16	DE	7 813 350	524 502	524 502
	EN	4 736 665	472 025	472 025
	FR	6 470 617	481 821	481 156
	RU	4 302 888	625 074	625 074

表3　3个语料库在3种分词粒度上的D值

语料库	语言	字符粒度	子词粒度	词粒度
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
News-Commentary-v16	DE	0.839	0.662	0.852 7
	EN	0.849	0.737	0.853 3
	FR	0.840	0.738	0.854
	RU	0.807	0.601	0.811

表4　3个语料库在3种分词粒度上的$F_{95\%}$值

语料库	语言	字符粒度	子词粒度	词粒度
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
News-Commentary-v16	DE	1	10	1
	EN	1	4	1
	FR	1	5	1
	RU	1	18	1

大语料库A的DTD值,即增大语符不平衡度. 语料库A在词粒度上的D值为$D_1 = D[1,2,3,4,5] = 0.40$,在子词粒度上的$D$值为$D_2 = D[1,5,6,8] = 0.40$. 结果表明语料库A子词粒度和词粒度具有相同的$D$值,即两者的语符不平衡度相同. 然后,本文将语料库A的子词粒度和词粒度的语符按频数从小到大进行排序并绘制在平面坐标系中,如图3所示. 当语料库中的所有语符频数相同时,坐标系中的点分布在一条水平线上,此时语料库的语符不平衡度为0. 语符分布线越倾斜(整体走势越陡),表明语符不平衡度越大. 通过观察图3可以发现,语料库A子词粒度的语符分布线比词粒度走势更陡,说明子词粒度具有更大的语符不平衡度. 由此可知,DTD算法相比D算法能更准确地衡量语料库的语符不平衡度.

4种算法在3个语料库上的运行时间如表6所示. 观察表6的数据可以发现,4种算法的时间复杂度具有以下关系:D > DTD >ρ> $F_{95\%}$. 4种算法的时间复杂度都不高,最长运行时间也仅为1.625 s,并且本文提出的DTD算法相比D算法时间复杂度降低了一个数量级.

通过以上对4种算法的对比分析可知,DTD算法相比D算法降低了时间复杂度,提升了准确度和鲁棒性;相比ρ和$F_{95\%}$算法,DTD算法能更全面、更有效地反映

表5 3个语料库在3种分词粒度上的DTD值

语料库	语言	字符粒度	子词粒度	词粒度
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
News-Commentary-v16	DE	7.32E5	4.92E3	1.84E3
	EN	6.16E5	5.23E3	2.90E3
	FR	7.96E5	5.70E3	3.10E3
	RU	5.73E5	4.57E3	1.69E3

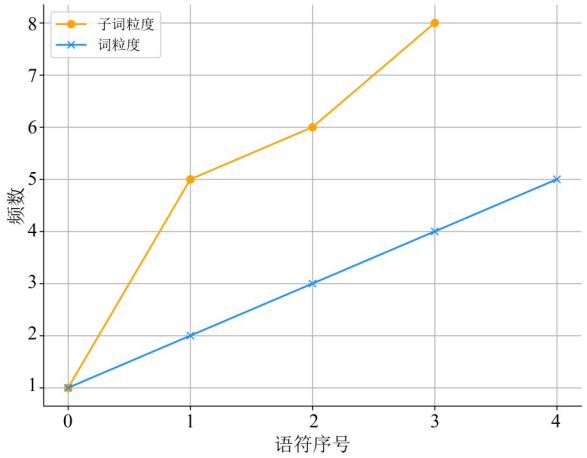


图3 语料库A的子词粒度和词粒度的语符分布

表6 算法ρ、D、$F_{95\%}$和DTD在3种语料库上的运行时间 单位：s

语料库	ρ	D	$F_{95\%}$	DTD
News-Commentary-v10	0.016	0.300	1.24E−05	0.062
Common Crawl corpus	0.156	1.625	1.16E−05	0.344
News-Commentary-v16	0.016	0.219	8.50E−06	0.031

语料库的语符不平衡度.

4.2 分词粒度对比分析

通过观察表5中的数据可知,语符不平衡度在分词粒度上具有以下规律:字符粒度>子词粒度>词粒度. 然而,仅仅通过观察图2中的语符分布很难找出这种规律背后的原因,所以本文结合表1中的关键参数进行了分析. 通过观察表1中的数据可知,相比于子词粒度,字符粒度的词表规模虽然大幅减小,但是语符频数的最大值Max$[C_i]$却显著增大. 例如,News-Commentary-v10语料库的FR语言子词粒度的词表大小是29 362,字符粒度的词表大小是206,降低了两个数量级;子词粒度的最大语符频数是292 188,字符粒度的最大语符频数是6 098 699,提高了一个数量级. 此外,通过观察表1中的K值可以发现,语料库经过字符级分词后频数为1的语符占比为10%~30%(相较于子词粒度,尽管比例K有所提升,数量N^*却显著减少). 因此,相较于子词粒度,字符粒度的语符分布线的整体走势更陡(分布线横向跨度显著变小,最大值显著变大,最小值不变),因而语符不平衡度更大.

语料库A的实验结果已经表明,子词级分词相较于词级分词会使语符分布线的整体走势变陡,从而增大语料库的语符不平衡度. 同样地,通过分析表1中的数据可以再次验证这一结论. 表1中,子词级分词相比于词级分词,词表规模显著减小了,语符的最大频数几乎不变(News-Commentary-v10的FR语料库变化最大,仅增大了0.156%),也依然存在一些低频语符,且低频语符比例和数量都显著降低了,所以其语符分布线的走势更陡(分布线横向跨度显著变小,最大值几乎不变,最小值不变),语符不平衡度更大. 子词级分词相比词级分词,除了会增大语料库的语符不平衡度,还会降低一些低频词的翻译效果. 例如,对于语料库A,子词级分词虽然会提升低频词“taller”和“cheaper”的翻译效果,却会使另一低频词“desk”相对其他语符的频数比更低,训练会更加不足,翻译效果也就更差. 但是,子词级分词相比于词级分词也有其优点所在. 如表1的K值所示,通过词级分词,词表中频数为1的语符占比为40%~60%,这意味着语料库中存在很大比例的低频语符,子词级分词可以将这一比例降至2.5%以下. 通过使用子词,词表大小以及低频语符的比例都有效减小,即使部分低频语符的翻译效果会受到影响,但整体翻译性能得到提升.

5 DTD算法验证

本文通过测试语符不平衡对NMT模型翻译效果的影响来验证DTD算法在表示语料库语符不平衡度方面

的准确性. 语符不平衡会造成高频语符过拟合,低频语符训练不足. 由于目前NMT模型训练有防过拟合措施,所以语符不平衡的主要不利影响是低频语符欠拟合. 于是,本文提出TRS(Token Rarity of Sentence)算法来计算语句的语符平均稀有度,算法如式(6)所示. 其中,l表示句子的语符个数,p_i表示每个语符在语料库中的频率. 利用TRS算法对训练语句进行排序,选取TRS值最小的10%语句构建测试集,用相应语料训练的NMT模型进行翻译,并计算Bleu得分. 理论上,当采用不同分词方法处理语料库时,语符不平衡度越大,Bleu得分越低.

$$\mathrm{TRS}=\frac{1}{l}\sum_{i=1}^{l}p_i \tag{6}$$

实验选取News-Commentary-v16的DE-EN,FR-EN和RU-EN 3个平行语料库作为数据集. 并且,3个平行语料库均选择EN作为目标语料库,以避免不同语言对实验结果产生影响. 在对源语料库和目标语料库进行预处理后,对语料库分别进行字符级、子词级和词级的分词,并计算相应的DTD值,结果如表7所示. NMT模型选择Facebook发布的Fairseq作为基本架构,Adam作为优化器,训练的相关参数设置为:adam-betas = (0.9, 0.98), clip-norm = 0.1, lr = 0.000 1, lr-scheduler = inverse_sqrt, weight-decay = 0.000 1, min-lr = 1e-09, dropout = 0.1, warmup-init-lr = 10^{-7}, warmup-updates = 4 000, criterion = label_smoothed_cross_entropy, label-smoothing = 0.1. NMT模型在3种分词粒度上的3个平行语料库训练完成后,使用Moses的multi-bleu计算模型在测试集上翻译结果的Bleu分数,结果如表8所示. 由于字符级的翻译结果转换成词级需要人工完成,为了方便对比,子词级和词级的翻译结果均转换成字符级后再进行Bleu分数的计算.

表7 News-Commentary-v16的3个平行语料库在3种分词粒度上的DTD值

平行语料库	字符级	子词级	词级
DE-EN	6.16E5	5.23E3	2.90E3
FR-EN	5.86E5	4.97E3	2.79E3
RU-EN	5.70E5	4.70E3	2.66E3

表8 News-Commentary-v16的3个平行语料库在3种分词粒度上的Bleu得分

平行语料库	字符级	子词级	词级
DE-EN	67.67	68.70	70.81
FR-EN	64.72	67.26	67.55
RU-EN	60.92	61.51	63.53

DTD算法表明语料库在3种分词粒度上的语符不平衡度规律为:字符粒度>子词粒度>词粒度. 表8显示3种分词粒度上训练的NMT模型相应的Bleu得分具有以下规律:字符粒度<子词粒度<词粒度. 这说明对于语料库而言,分词粒度不同时,语符不平衡度越大,低频语符的翻译效果越差,也验证了DTD算法表示语料库语符不平衡度的正确性.

6 结论

本文提出了一种新的算法,用以评估NMT语料库的语符不平衡度,并利用该算法分析不同语料库在字符、子词和词3种粒度上的语符不平衡情况. 本文的贡献如下:

(1)提出DTD算法,用以计算语符不平衡度. 该算法在准确度、有效性和鲁棒性方面相比之前的研究有较大提升;实验结果在不同分词粒度间的差异更加显著,便于对比分析.

(2)将分词范围从以往研究的子词粒度扩展到NMT最常用的3种分词粒度——字符、子词和词,并发现字符粒度的语符不平衡度最大,子词粒度次之,词粒度最小.

(3)进行DE,EN,FR,RU这4种语言的对比实验,发现了这4种语言在语符不平衡度上的规律性和差异性.

参考文献

[1] CHO K, VAN MERRIENBOER B, GULCEHRE C, et al. Learning phrase representations using RNN encoder-decoder for statistical machine translation[C]//Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). Stroudsburg: Association for Computational Linguistics, 2014: 1724-1734.

[2] SUTSKEVER I, VINYALS O, LE Q V. Sequence to sequence learning with neural networks[C]//Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2. New York: ACM, 2014: 3104-3112.

[3] BAHDANAU D, CHO K, BENGIO Y. Neural machine translation by jointly learning to align and translate[EB/OL]. (2014)[2021]. https://arxiv.org/abs/1409.0473.

[4] PAPINENI K, ROUKOS S, WARD T, et al. BLEU: a method for automatic evaluation of machine translation [C]//Proceedings of the 40th Annual Meeting on Association for Computational Linguistics - ACL'02. Morristown: Association for Computational Linguistics, 2001: 311-318.

[5] SENNRICH R, HADDOW B, BIRCH A. Neural machine

translation of rare words with subword units[C]//Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Stroudsburg: Association for Computational Linguistics, 2016: 1715-1725.

[6] KOEHN P, OCH F J, MARCU D. Statistical phrase-based translation[C]//Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology - NAACL'03. Morristown: Association for Computational Linguistics, 2003: 127-133.

[7] CHIANG D. Hierarchical phrase-based translation[J]. Computational Linguistics, 2007, 33(2): 201-228.

[8] LUONG M T, MANNING C D. Achieving open vocabulary neural machine translation with hybrid word-character models[C]//Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Stroudsburg: Association for Computational Linguistics, 2016: 1054-1063.

[9] WU Y H, SCHUSTER M, CHEN Z F, et al. Google´s neural machine translation system: Bridging the gap between human and machine translation[EB/OL]. (2016) [2021]. https://arxiv.org/abs/1609.08144.

[10] LEE J, CHO K, HOFMANN T. Fully character-level neural machine translation without explicit segmentation[J]. Transactions of the Association for Computational Linguistics, 2017, 5: 365-378.

[11] CHERRY C, FOSTER G, BAPNA A, et al. Revisiting character-based neural machine translation with capacity and compression[C]//Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. Stroudsburg: Association for Computational Linguistics, 2018: 4295-4305.

[12] GOWDA T, MAY J. Finding the optimal vocabulary size for neural machine translation[C]//Findings of the Association for Computational Linguistics: EMNLP 2020. Stroudsburg: Association for Computational Linguistics, 2020: 3955-3964.

[13] MIKOLOV T, CHEN K, CORRADO G, et al. Efficient estimation of word representations in vector space[EB/OL]. (2013)[2021]. https://arxiv.org/abs/1301.3781.

[14] 尤洪峰, 田生伟, 禹龙, 等. 基于Word Embedding的遥感影像检测分割[J]. 电子学报, 2020, 48(1): 75-83.
YOU H F, TIAN S W, YU L, et al. Remote sensing image detection and segmentation based on word embedding [J]. Acta Electronica Sinica, 2020, 48(1): 75-83. (in Chinese)

[15] 尤洪峰, 田生伟, 禹龙, 等. 基于Word Embedding的遥感影像检测分割[J]. 电子学报, 2020, 48(1): 75-83.
YOU H F, TIAN S W, YU L, et al. Remote sensing image detection and segmentation based on word embedding [J]. Acta Electronica Sinica, 2020, 48(1): 75-83. (in Chinese)

[16] 张昱, 刘开峰, 张全新, 等. 基于组合-卷积神经网络的中文新闻文本分类[J]. 电子学报, 2021, 49(6): 1059-1067.
ZHANG Y, LIU K F, ZHANG Q X, et al. A combined-convolutional neural network for Chinese news text classification[J]. Acta Electronica Sinica, 2021, 49(6): 1059-1067. (in Chinese)

[17] 李维刚, 甘平, 谢璐, 等. 基于样本对元学习的小样本图像分类方法[J]. 电子学报, 2022, 50(2): 295-304.
LI W G, GAN P, XIE L, et al. A few-shot image classification method by pairwise-based meta learning[J]. Acta Electronica Sinica, 2022, 50(2): 295-304. (in Chinese)

[18] 胡峰, 王蕾, 周耀. 基于三支决策的不平衡数据过采样方法[J]. 电子学报, 2018, 46(1): 135-144.
HU F, WANG L, ZHOU Y. An oversampling method for imbalance data based on three-way decision model[J]. Acta Electronica Sinica, 2018, 46(1): 135-144. (in Chinese)

[19] 徐婕, 贺美美. 基于马氏抽样的SVM非平衡数据分类算法的泛化性能研究[J]. 电子学报, 2018, 46(11): 2660-2670.
XU J, HE M M. Research on the generalization performance of SVM imbalanced data classification algorithm based on Markov sampling[J]. Acta Electronica Sinica, 2018, 46(11): 2660-2670. (in Chinese)

[20] 陆克中, 陈超凡, 蔡桓, 等. 面向概念漂移和类不平衡数据流的在线分类算法[J]. 电子学报, 2022, 50(3): 585-597.
LU K Z, CHEN C F, CAI H, et al. Online classification algorithm for concept drift and class imbalance data stream[J]. Acta Electronica Sinica, 2022, 50(3): 585-597. (in Chinese)

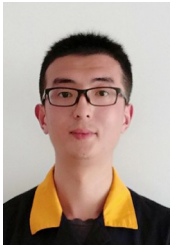
[21] JOHNSON J M, KHOSHGOFTAAR T M. Survey on deep learning with class imbalance[J]. Journal of Big Data, 2019, 6(1): 1-54.

[22] RUBNER Y, TOMASI C, GUIBAS L J. The earth mover's distance as a metric for image retrieval[J]. International Journal of Computer Vision, 2000, 40(2): 99-121.

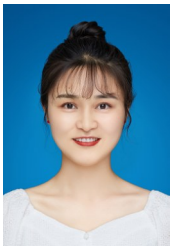
[23] CHAO Y R, ZIPF G K. Human behavior and the princi-

ple of least effort: An introduction to human ecology[J]. Language, 1950, 26(3): 394.

作者简介



王海波 男,1995年生,安徽合肥人. 浙江大学生物医学工程与仪器科学学院硕士研究生. 主要研究方向为自然语言处理、机器翻译、人工智能.
E-mail: wanghaibo111@zju.edu.cn



余丽丽 女,1994年生,安徽合肥人. 浙江师范大学教师教育学院博士研究生. 主要研究方向为语言学、多语种翻译、自然语言处理.
E-mail: yulili@zjnu.edu.cn

王宏伟(通讯作者) 男,1981年生,黑龙江齐齐哈尔人. 于英国剑桥大学获博士学位. 现为浙江大学伊利诺伊大学厄巴纳香槟校区联合学院长聘教授、博士生导师. 主要研究方向为人工智能、知识图谱、工业大数据以及故障诊断.
E-mail: hongweiwang@zju.edu.cn