arXiv:2303.02427v2 [cs.CL] 4 Apr 2023

无监督跨语言分词的自调超参数

Anton Kolonin[1], 2[0000−0003−4180−2870]

[1] 俄罗斯新西伯利亚国立大学，新西伯利亚，Pirogova 1，630090，俄罗斯 https://nsu.ru/[2] Aigents，俄罗斯新西伯利亚，Pravdy 6-12，630090，俄罗斯 https://aigents.com/ akolonin@gmail.com

摘要。 我们探讨了针对英语、俄语和中文的语言无关无监督分词问题进行元学习的可能性。我们实现了元学习方法，用于自动确定先前研究提出的无监督分词模型的超参数；该方法依赖于多种独立于人工的适应度函数，例如归一化反熵、压缩因子和交叉分割 $F1$ 得分，以及这三项指标的加法和乘法复合组合，并将它们与传统的 $F1$分词得分进行比较。我们发现，对于英语和俄语，后者与前三项指标的加法组合之间存在相当强的相关性。对于中文，我们发现 $F1$ 得分与压缩因子之间存在显著相关性。我们的结果表明，对低资源语言和已消亡语言进行稳健的无监督分词是可行的，也使我们能够从不同人类文化演化出的不同结构优化方案出发，将人类语言视为高效符号通信编码的演化产物。

关键词： 跨语言 · 语言学习 · 元学习 · 分词 · 无监督。

1 引言

早期依赖基于可解释图的模型，以及依赖概率指标和表达不确定性的指标（如“转移自由度”或“转移自由度”）开展无监督分词学习的研究已显示出良好成果，尤其是将后类指标应用于英语、俄语和中文时 [1,2,3]。

然而，此前研究人员针对英语、俄语和中文开展的最新跨语言研究 [3]发现，为了获得高度准确的分词输出，不同语言在模型构建及其后续应用过程中需要使用不同的超参数。这意味着，为了获得一种完全无监督的语言

习得任务解决方案，需要从分词层面入手，借助一些额外的指标来驱动特定语言模型的构建，并确定这些模型在应用时所需的参数，且无需人工监督或有关该语言本身的先验知识。

为了解决后一个问题，本文参照近期的相关研究 [3]，探索在语言无关的无监督分词设置中采用元学习的可能性。我们实现了元学习方法，用于自动确定应用于未标注的原始训练文本语料库的超参数，并搜索独立于人工的适应度函数，例如归一化反熵、压缩因子和所谓的“跨切分 $F1$ 得分”，以及这三项指标的加法与乘法组合，并将它们与传统的 $F1$ 分词得分进行比较。

2 实验方法

2.1 超参数调优指标

为了找到一种独立于人工且不受文化影响的指标，以优化人类语言等涉及大规模人际交流的符号序列的切分，我们从几个不同角度着手。第一是“压缩”视角，即从人类语言具有信息压缩特性的角度出发 [4], ，该观点认为，每种人类语言的语言结构都在文化演化过程中以实现信息压缩为目标。第二是“信息”视角，即从信息论的角度出发，认为语言结构可能由一个客观目标驱动：在该语言传递的所有可能消息中实现熵最小化 [5]。第三是“一致性”视角，即从语言的社会契约角度出发，将语言视为一种媒介，使不同群体之间能够有效沟通，并尽量减少这些群体间社会语境的差异 [6,7]。

事实上，上述三种视角至少可以看作互补，甚至可以说是等价的。例如，衡量最优压缩的指标，可以从使熵最小化的角度来看，也可以从相应的度量角度来看 [8] 。此外，正如我们将在研究中看到的，我们发现，语言模型在不同训练语料（这些语料可归因于不同社会群体）之间提供共识的“折中”能力，也与前述“压缩”和“信息量”指标高度相关。下面我们定义这三种度量。

我们将“压缩”分词指标称为“压缩因子”$C\%$，它对应于一个比值：分子是使用当前分词模型对训练集进行压缩后的大小，分母是训练集未压缩时的大小。压缩后的大小通过以下两部分计算得出：分词后的文本语料库中词元索引序列的长度，以及“词典”的大小——其中所有词元长度之和。训练集未压缩时的大小则仅按其中的字符数计算。

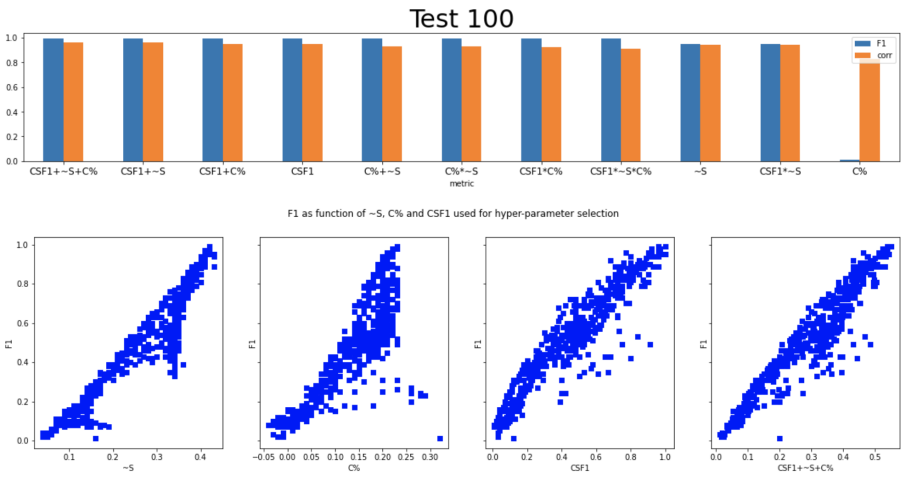

图 1。 针对英语和包含 100 行的测试集。上排：蓝色条形表示基于与横轴上目标指标最大值对应的超参数组所获得的 $F1$ 分数；蓝色条形表示 $F1$ 分数与相应指标之间的皮尔逊相关系数。

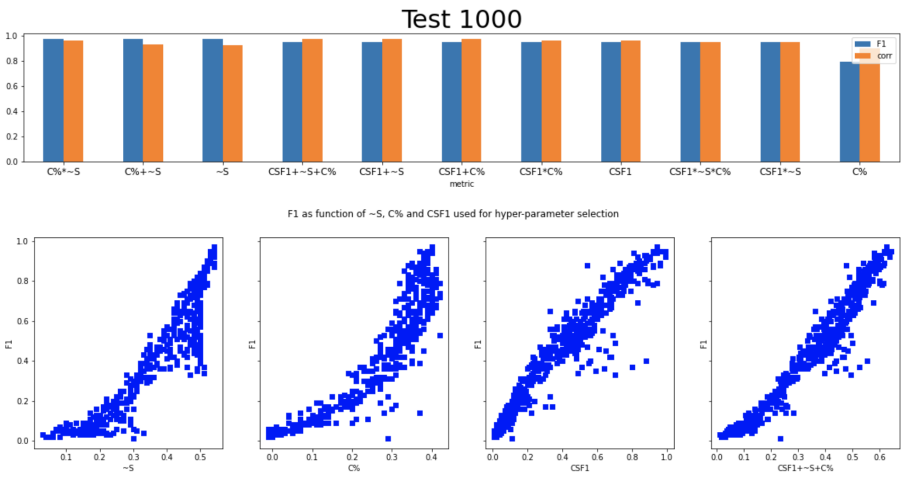

图 2。 针对英语和包含 1000 行的测试集。上排：蓝色条形表示基于与横轴上目标指标最大值对应的超参数组所获得的 $F1$ 分数；蓝色条形表示 $F1$ 分数与相应指标之间的皮尔逊相关系数。

对于“信息”分词指标，我们引入了所谓的“归一化˜ ˜反熵” S，其定义为 $S = 1 - H/(log_2(L))$，其中 H 是一个香农

使用给定分词模型对整个训练集进行分词后，其熵以 2 为底； L 是该分词模型下词库的大小。

对于“一致性”指标，我们采用一种称为跨分割 $F1$ 得分的指标，其定义如下。首先，将训练集语料库分成两部分，大小相同，分别称为集合 A 和集合 B。接着，根据前文引用的工作 [3] ，分别为每部分语料库创建跨 N 元语法的图遍历模型，将它们记为 M(*A*)和 M(*B*)。然后，使用这两个模型对测试集进行分词，从而得到 T(*M*(*A*))和 T(*M*(*B*))。最后，将其中一个作为另一个的“真实值”，评估分词的跨分割 $F1$得分：以 $CSF1$为 T(*M*(*A*))，对比 T(*M*(*A*))。

2.2 调优超参数

实验设置包括：针对参考文献中所使用的测试集，寻找用于分词的传统 $F1$ score 指标 [3], ；同时，使用同一文献中的“无转移自由度”作为驱动分词的指标，并考察如下三个超参数构成的三维网格：

– N 用于执行模型图遍历的秩组合，以及基于指定 N 元语法子集计算的“均值”指标。我们探索了 N 在 1 到 7 范围内的每个可能取值（中文情况下为 1 到 3），以及 N值的各种组合。– 模型压缩阈值 T_{mc} 用于移除低频 N 元语法（在模型图中对应顶点及其之间的转移）。我们使用了以下取值：0.0（完全不进行压缩）、0.0001、0.001、0.01 和 0.1。– 分词指标阈值 T_{tm}：指标值超过此水平时，即对应一个词元边界。我们使用了以下取值：0.0001、0.0005、0.001、0.005、0.01、0.02、0.05、0.1、0.2、0.3、0.4、0.5、0.6、0.7、0.8 和 0.9。

除3维超参数空间中每个点的 $F1$ 得分外，还使用相同的超参数值评估了目标指标 C%、 S 和 $CSF1$ 。我们还计算并评估了˜ ˜基于这三个主要指标的复合指标，例如加法 C%+S 或 C%+S+$CSF1$ ，以及˜ ˜乘法 C%∗S 或 C%∗S∗$CSF1$。

2.3 训练集与模型构建

为进行模型构建，根据所引用的工作 [3], ，我们使用了公开的英语、俄语和中文语料库——与前一篇论文相同。

对于英语，我们使用了Brown训练语料库（大小为6M字节），该语料库从http://www.sls.hawaii.edu/bley-vroman/brown$ $nolines.t xt下载。

对于俄语，我们使用了从 https://www.kaggle.com/datasets/oldaandozerskaya/fiction-corpus-for-agebased 下载的 RusAge语料库

-text-classification，其中最大的语料库名为 Previews（大小为 825M 字节），用于构建模型。

对于中文，我们使用了从 https://github.com/brightmart/nlp chinese corpus 下载的 CLUE基准新闻2016数据集。下载后发现，文件夹 new2016zh 中有两个文件：news2016zh valid.json 和 news2016zh train.json。随后，我们对较大的文件（大小为 8930M 字节）进行了程序化处理（解析 JSON；选取 title、desc 和 content 字段；并将每个字段分别保存为独立的行），生成了纯文本文件，并将其用于模型构建。

模型构建完全按照参考文献 [3] 中所述的相同流程进行，并为每种语言分别得到三个模型。

2.4 测试集

为了确定 $F1$ 得分和 $CSF1$ 指标，每种语言都使用了两个测试集。

首先，有一个包含 100 行文本的测试集，这些文本来自一个由 100 条与个人理财相关的多句陈述组成的中英平行语料库。该语料库以制表符分隔文本文件的形式从 Magic Data (https://magichub.com/datasets/chinese-english-parallel-corpus-finance) 下载，其中分别设有名为 zh 和 en 的列，存放中文和英语版本。该语料库还扩充了俄语翻译，只有一列名为 ru，其中包含俄语翻译；这些内容可在 Aigents/Pygents开源项目中的文件 (https://github.com/aigents/pygents/blob/main/data/corpora/Russian/magicdata/zh_en_ru_100/CORPUS_ZH_EN_RU.txt) 中找到。

2.5 实验方法摘要

也就是说，我们获得了函数 $F1(N_s, T_{mc}, T_{mc})$、$C\%(N_s, T_{mc}, T_{mc})$、$\sim S(N_s, T_{mc}, T_{mc})$ 和 $CSF1(N_s, T_{mc}, T_{mc})$，并针对英语、俄语和中文，基于所引用文献 [3], 的附录及参考文献中 Aigents/Pygents开源项目（ https://github.com/aigents/pygents/）中的代码所描述的模型和流水线，在包含 100 行和 1000 行文本的测试集上探索了它们之间的相关性和对应关系。

我们使用后面三个目标指标及其导数来寻找与之对应的超参数组合，然后利用这些超参数执行“预期最佳”的分词，计算 $F1$ 得分，并将其值与该指标关联起来。

与此同时，我们使用超参数三维空间中的所有点，采用这些参数对测试集进行分词，并计算此类分词的 $F1$ 分数与基于这些分词计算出的相应目标指标和复合指标之间的皮尔逊相关系数值。

3 实验结果

表1展示了在所探索的超参数三维空间中的所有组合下，各目标指标两两之间的皮尔逊相关系数。表2展示了目标指标与基于这些指标构建的部分复合指标之间的皮尔逊相关系数。

可以看出，根据表1，$\tilde{S}$ 与 $C\%$ 和 $CSF1$ 呈显著正相关。此外，对于英语和俄语，所有指标之间都高度相关，其中 S 与另外两个指标的相关性尤其高。相比之下，对于中文，所有相关性普遍较低，而 $CSF1$ 与其他任何指标都不相关。

此外，表2显示，$F1$ 得分与复合 $C\%+S$ 的相关性最高；至于其他指标，它们在某些语言中的表现更好，在另一些语言中则更差，详见下文按语言分类的章节。

语言	指标1	指标2	相关性
英语	$CSF1$	$\tilde{S}$	$\mathbf{0.90}$
英语	$CSF1$	$C\%$	$\mathbf{0.86}$
英语	$\tilde{S}$	$C\%$	$\mathbf{0.92}$
俄语	$CSF1$	$\tilde{S}$	$\mathbf{0.76}$
俄语	$CSF1$	$C\%$	$\mathbf{0.64}$
俄语	$\tilde{S}$	$C\%$	$\mathbf{0.83}$
中文	$CSF1$	$\tilde{S}$	$\mathbf{0.36}$
中文	$CSF1$	$C\%$	$\mathbf{-0.31}$
中文	$\tilde{S}$	$C\%$	$\mathbf{0.48}$

表1。基于包含1000行的较大测试集，不同语言的目标指标之间的皮尔逊相关系数。

3.1 英语

对于英语，我们发现，几乎可以基于任何具有 $CSF1$、S 的指标来自动调节超参数，而它们的组合指标尤其适合这一目的，参见表2、图1和图2。

需要注意的是，图1和图2中对应于截然不同的测试集的点分布非常接近，可以将其视为描述特定语言模型的“特征”。

3.2 俄语

对于俄语，我们发现这种方法的可靠性明显较低，但仍然可行；它依赖于 $CSF1$、S 及其组合指标（与英语的情况相同），详见表2、图 3和图 4。

语言指标相关性		
英语	$CSF1$	**0.95**
英语	$C\%$	**0.83**
英语	$\tilde{S}_{_C\%}$	**0.94**
英语	$+S_{_C\%*S}$	**0.93**
英语		**0.93**
俄语	$CSF1$	**0.79**
俄语	$C\%$	**0.56**
俄语	$\tilde{S}_{_C\%}$	**0.87**
俄语	$+S_{_C\%*S}$	**0.67**
俄语		**0.68**
中文	$CSF1$	$\mathbf{-0.09}$
中文	$C\%$	**0.85**
中文	$\tilde{S}_{_C\%}$	**0.79**
中文	$+S_{_C\%*S}$	**0.95**
中文		**0.83**

表2. 基于包含1000行的更大测试集，展示不同语言的各项指标与分词 $F1$ 得分之间的皮尔逊相关系数。

与英语的情况相同，图 3和图 4中对应不同测试集的点分布仍然相当相似，似乎可以作为语言模型的“签名”。

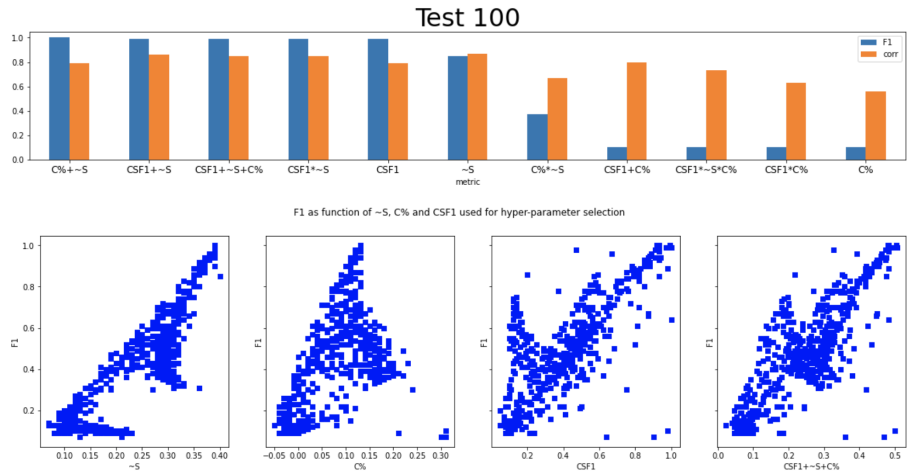

图 3. 针对俄语以及包含 100 行的测试集。上排：蓝色柱表示基于与横轴上最大目标指标相对应的超参数集所获得的 $F1$ 分数，蓝色柱表示 $F1$ 分数与相应指标之间的皮尔逊相关系数。



图4. 对于俄语和包含1000行的测试集。上排：蓝色柱表示基于横轴所示目标指标达到最大值时对应的超参数集合计算出的 $F1$ 分数；蓝色柱表示 $F1$ 分数与相应指标之间的皮尔逊相关系数。

3.3 中文

对于中文，结果不如英语和俄语明确。不过，˜将 C%与 S 结合起来，仍可实现超参数的自动确定，参见表2和图5。

需要指出的是，使用包含100行的测试集完全无法为中文提供足够合理的结果，因此本文未予展示。

4 结论

在全无监督分词的超参数空间中，对原始自然语言文本进行分词时，我们发现，针对英语和俄语，所有三种已引入的与人类无关的指标——“压缩因子”“归一化反熵”以及所谓的“交叉拆分 $F1$ 分数”——彼此之间具有相当强的皮尔逊相关性。我们还发现，这些参数及其组合指标与使用相应超参数集执行的无监督分词过程的准确率之间具有很强的皮尔逊相关性。

对于中文，我们发现，$F1$分数与“压缩因子”和“归一化反熵”之间存在显著相关性，这两者的组合指标也同样如此。

我们的结果表明，有可能对低资源语言和已消亡语言进行稳健的无监督分词，并使我们能够从以下角度思考人类语言：

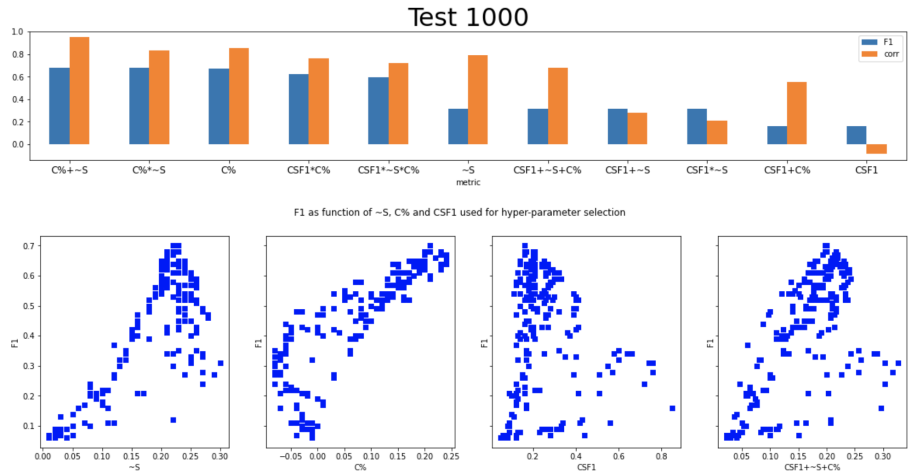

图 5. 对于中文语言和包含 1000 行的测试集。上排：蓝色柱表示根据横轴所示目标指标达到最大值时对应的超参数集合所获得的 $F1$ score；蓝色柱表示 $F1$ score 与相应指标之间的皮尔逊相关系数。

效率高的符号通信编码的演化，以及不同人类文化中演化出的不同结构优化方案。

与此同时，这也为基于未标注且未分词语料库进行端到端无监督语言学习开辟了道路，正如 Vepstas 和 Goertzel [9] 所提出的，并由 Glushchenko 等人进一步发展 [10]。

局限性

这项工作的主要局限似乎在于，所用的三种语言语料库在规模和主题方面差异很大。英语和俄语语料库均为文学作品，而中文语料库则基于爬取的新闻数据。此外，英语语料库的规模只有俄语语料库的百分之一，而中文语料库的规模则是后者的十倍。这些因素中的任何一个，或二者共同作用，都可能解释为何英语的结果优于俄语，以及俄语的结果优于英语。

在后续工作中，我们计划使用基于平行多语语料库构建的模型，重新进行同一项实验；这些语料库在所有参与实验的语言中主题相同、规模一致。

进一步研究的另一个方向，可以是针对每种语言，使用规模不同的训练语料库开展相同的实验，以了解语料库的规模是否也能成为推动超参数自调性能的另一个超参数。

伦理声明

这项工作没有明显的负面伦理影响。相反，我们认为，这项工作具有积极的伦理价值，因为它促进了对所谓“低资源”语言和方言文化的包容，而这些语言和方言难以通过当前的语言学方法进行研究。我们相信，所提出的技术可以通过基于原始田野数据提供初始词汇表，帮助人们学习这类语言，从而为进一步学习这些语言及其语法铺平道路。这项工作的另一项长期积极伦理影响，源于其“可解释”的特性。我们的模型有助于转向开放、透明且人性化的语言模型，这类模型可针对任何人类语言开发并投入生产，从而避免“黑箱”自然语言处理模型可能导致生活质量下降。我们预计这项工作不会造成有害的社会影响。

参考文献

1. Logan Kearsley. 跨语言分词的混合方法：统计形态学。载于杨百翰大学学位论文与学位论文集，第5984号。2016。 2. Jesse O. Wrenn、Peter D. Stetson 和 Stephen B. Johnson. 一种无监督机器学习方法，用于分割临床医生输入的自由文本。载于 AMIA年度研讨会论文集，第811–5页。2007。 3. Anton Kolonin 和 Vignav Ramesh. 无监督分词学习。载于2022年自然语言处理实证方法会议论文集，第3649–3664页。2022。 4. Simon Kirby、Monica Tamariz、Hannah Cornish 和 Kenny Smith. 语言结构文化演化中的压缩与交流。Cognition，141：87–102。2015。 5. Eugene Kharitonov、Rahma Chaabouni、Diane Bouchacourt 和 Marco Baroni. 涌现语言中的熵最小化。arXiv:1905.13687 [cs.CL]. 2019。 6. John R. Searle. 语言作为生物与社会现象。载于《构建社会世界：人类文明的结构》。牛津大学出版社。2010。 7. Abdurrahman Usman、Adi Mahmud、Abdulhalim Daud 和 Suratman Dahlan. 语言作为一种社会工具。Edukasi，18：259–276。2022。 8. Georges Hansel、Dominique Perrin 和 Imre Simon. 压缩与熵。载于 STACS 92，第513–528页，柏林、海德堡。Springer Berlin Heidelberg。1992。 9. Linas Vepstas 和 Ben Goertzel. 从大型（未标注）语料库中学习语言。2014。arXiv:1401.3372 [cs.CL]. 2014。 10. Alex Glushchenko、Andres Suarez、Anton Kolonin、Ben Goertzel 和 Oleg Baskov. 面向无监督语言学习的程序化链接语法归纳。载于人工通用智能，第111–120页，Cham。Springer International Publishing。2019。