Nihao: Relative Fine-Grained Chinese Word Segmentation Criterion

Kaiyu Huang, Hao Yu, Wei Liu, Degen Huang*
School of Computer Science, Dalian University of Technology

April 2021

摘　要

本文为Nihao分词标准–中文版，为研究人员对Nihao分词数据集及其切分准则提供更好的理解，促进汉语分词的应用和更深层的探索。

1　责任声明

本标准仅用于研究目的，文中内容不涉及政治和隐私性敏感话题，引用与本标准相关信息务必引用本文。
本标准一切说明及解释权由大连理工大学自然语言处理研究室（DUTNLP Lab）所有。
组长：黄德根教授
成员：黄锴宇，余浩，刘伟

2　绪论

文档主要描述针对汉语文本的分词标准，附加少量英文文本及其他特殊字符的处理准则。Nihao分词主要面向信息检索与机器翻译的下游高层自然语言处理任务，因此，所提出标准为相对**细**粒度（fine-grained）分词标准，即，尽可能划分最小语义单元，避免出现多个语义单元组成的长词。切分规范主要作为规定现代汉语的切词原则，即什么样的汉字组合可以为一个切分单位。本分词标准主要以缩减句长，减少歧义，便于理解为目的。

3　基本概念

“切分单位”，是指信息处理中使用的、具有确定的语义和语法功能的基本单位。Nihao切分规范中的“切分单位”主要是词，也包括了一部分结合紧密、使用稳定的词组。在某些特殊情况下孤立的语素或非语素字也可能出现在切分

* Corresponding author

序列中[1]。规范规定，凡收入**词典**的词条（包括：词、习惯用语、简略语等）一**般**都是切分单位。“词典词条”（或“词条”）指在词典（这里指商务印书馆的《应用汉语词典》）中收录的那些词语。对于含有前缀后缀的词汇，按照修饰词表软性划分。从字数考虑,对两个字的组合较宽的看作是一个切分单位，三个字的较严格，四个字以上的若不是成语或习语一般不看作是一个切分单位。

4 切分规则

切分规则为人工拟定规范。由于规范与规则模板相似，因此，规范间会不可避免地存在重叠及冲突情况。部分切分单位的实际划分与准则有一定出入，具体黄金切分一切以人工标注为准。如下游任务对领域术语有特殊需求，根据具体下游任务可通过术语词典进行后处理加工，本规则以通用领域为基准。

4.1 专有名词

4.1.1 人名

详见表1.

	组成结构	示例
普通人名全名	姓+名	张/路；欧阳/修; 李/思思
人名搭配	姓名简写+职务等 后缀修饰性名词	张/处长；小平/同志 陈/总；王/校长
	前缀修饰+姓名简 写（非专有名词）	老/张；小/张 阿/黄；阿/华
	带有明显排行修饰 词+人的简称、尊称等	二/婶 三/哥
笔名 少数民族人名 外国人名	整体成词	鲁迅 成吉思汗 卡尔·马克思

表 1: 人名切分标准

4.1.2 地名

详见表2.

4.1.3 组织机构名

总的原则：**以最小语义实体划分**。

[1]补充：切分原则应人为引入指导概念，切分是为了降低自然句理解难度，降低歧义。

	组成结构	示例
基本地名	单独成词	阿根廷；泰山；阿克苏 华山；鸭绿江
复合地名	拆分后每个单词都有意义，就将其拆分；否则合并。	中华/人民/共和国 美利坚/合众国；日本/国
	地名后接省，市，县，区，乡，镇，村，旗，州，都，府，道等单字的行政区划名称时，要与地名切分开。地名后的行政区划长度大于2，也将地名同行政区划名称切分，行政区名大于两字的一般也需要切分。	新泽西/州 辽宁/省 东京/都 喀喇沁/旗 宁夏/回族/自治/区 香港/特别/行政/区 深圳/特区 华盛顿/特区
	地名后接表示地形地貌/自然区划，若地貌名大于2，则与地名切开。若地貌名为单字，则不予切分。	华北/平原 陶然亭/公园 中关村 沈阳路
缩略地名	最长片段缩略语单独成词	中日韩；黑吉辽

表 2: 地名切分标准

(1) 粗粒度机构名称按照其组成单位进行切分。例：大连/医科/大学,大连/理工/大学。

(2) 分开后无意义的合并在一起作为机构切分词段。例：联合国，国务院。

(3) 机构名称的简略语不切分, 例：北大,清华,大工,世贸,军委,市委,省委,等。

4.1.4 其他专有名词

(1) 除人名、地名、团体、组织、机构名称以外的其他专名，后接修饰词。若词长大于2，则多数情况下应当切分，若为单音节语素则一般不切分，表示民族的“族”、表示语言的“语”，表示文字的“文”。例：汉语；维吾尔语；维吾尔/语言；满族；俄罗斯族；俄罗斯/族人；仡佬族；茅盾奖；诺贝尔奖；爱斯基摩人。

(2) 专有交通线路（或简称）,若是复合语,按组成部分切分。例：山手/线；津浦/路；京九/铁路。

(3) 历史事件,按组成部分切分。例：卢沟桥/事变；五/四/运动；七/七/事变；甲午/海战；戊戌/变法。

(4) 品名由牌+普通名词组成，按组成部分切分。例：康师傅/方便面；小米/12/手机；万通/XI/型/门锁；老干妈。

(5) 带有序号数字的专有名词一般不作为整体专有名词，按组成结构切分。例：2/号/国道；十一/届/三中全会。

(6) 书、报、杂志、文档、报告、协议、合同等名称中通常有书名号加以标识，不作为专有名词。由于这些名词长度通常较长，品名本身按常规处理切分。例：宁波/日报；中华/读书/报。

(7) 食谱上的菜名等名词，按最小语义单元进行切分。例：紫菜/鸡蛋汤；宫保肉丁；木樨肉；玫瑰/芝麻饼。

(8) 特殊活动，即使在引号内，也被看作普通名词，按粒度切分。例：庆/回归/公益/千万/行。

4.2 代词

单音节代词“本”、“每”、“各”、“诸”后接单音节名词时，和后接的单音节/多音节名词，应予以切分。例：每/人；每/家；本/地区；各/部门。后面接单字的量词要切开，后接”些”不作为量词,不予以切分。例：这/个；某/个；一些；这些；那些。

4.3 数词和数量词组

(1) 一般数词视为一个切分单位。例：一百二十三；120万；123.53；三分之二；20%。

(2) 数词+量词的形式，按结构切分。例：一/个，一/方/沃土。

(3) 序数词，“第”+中文数字或阿拉伯数字,视为一个切分单位。例：第一；第一百零一/个。

(4) 概数词，如，许多、少数、很多、不少、大量等表示数的概念，一般作为修饰名词或量词的成分出现，应予以切分。例：许多/人；很多/个；一些；这些；那些；二十/左右/个；非常/多/人。

(5) 数字加上单音节语素表示专有名词或独立语义时,视为一个切分单位,不予切分。例：八卦；五行；三角；两性；万岁。

(6) 数字+多音节词,一般切分,若整体作为常用语/口语化表示则不予以切分。例：一/方面；一/池塘，一揽子。

(7) 年代日月乾坤被看作为单独切分单位，时间表达式按照组成结构切分。例：西周；南北朝；牛/年；1999/年/6/月/20/日/2/点/14/分。

4.4 双音节述补结构

双音节动词+动词或动词+形容词构成的述补结构。这类结构的切分单位较难划分，但是此类结构的不同切分方法对下游任务影响较小。在本分词标准中，根据独立语义单元进行切分。

(1) 若拆开各是一个词，通常作为两个切分单位。例：走/到；调/好；坐/稳。

(2) 若拆开了，其中至少有一个是语素，通常就不切分，作为一个切分单位。例：形成；鼓动；说明；震动。

(3) 双音节的述补结构(1)情况的中间插入“得”或“不”一般应予切分，(2)情况的中间插入“得”或“不”一般不予切分。例：走/得/到；走/不/到；形得成；形不成。如果去掉“得”或“不”后，前后两个字不构成一个词的，则作为一个切分单位。例：来得及；来不及；对得起；对不起；说得过去；说不过去。

4.5 时间和方位词

如“前”、“后”、“底”等表示时间前后的一般与相邻语义字段合并。例：会前，会上，年底，年末，年初，至今，近年来，此前，其后。
如“下”、“上”、“外”、“里”、“出”、“入”等表示方位或指向性方向的单音节语素，一般与相邻语义字段关联，不予以切分。例：山上，网上，此外，法外，仇外，选出，提出，记入，存入，入境，入世。

4.6 四字及四字以上词语

可被切分位2+2格式的四字词，一般切分成两个二字切分单位组合的形式。例：总结/经验；贯彻/执行；调查/研究。对于像“生产资料”、“国民经济”等虽然作为一个词已收入词典，但分开后，对于检索更加有效，所以仍然要将其分开。如：生产/资料；国民/经济。原则上，对于复合词一般予以切分，分开后的部分如果不成词，则不予切分，作为一个切分单位。如：计件/工资。本标准同时将提供一份补充常用长词词表，词表内的词一般情况下作为一个切分单位，为**最优先级**单独切分单位，详见附录1。

4.7 连词

兼做介词和连词的词作为一个切分单位，如：和，跟，同，与，于，等单音节词。范例说明：列/于；取决/于；限/于；载/于。其他的多音节连词一般也作为一个切分单位，例：但是、尽管、即使、如果、因此，等。

4.8 虚词和助词

动态助词视为一个切分单位，如：着、了、过、矣、然，等。语态助词也视为一个切分单位，如：云云、等、等等、而外、而已、似的、与否、喽、而言、焉、为、乎、话、不可、被，等。其中，“的”分为多种用法，当作为助词时，视为一个切分单位，例：丰富/的/经验；我/的/杂志。当“的”用在句末表示肯定，也单独作为一个切分单位，“他/是/要/去/的”。但“的”作为非助词用法时，由具体语境决定切分单位，如：的士，目的，无的放矢，等。其他结构助词，作为一个切分单位，如：之、为止、说来、来说、的话，等。

4.9 组合性复合词

汉语中存在大量由多个单独语义组合形成的符合词，本小节将对此类词的切分标准进行概述。

(1) “是、能、有、要、可”+单音节/多音节语素形成复合词，一般予以切分。例：不/是；不/能；是/不/是；能/不/能；不/要；就/要；就/是；还/要；既/有；应/有；有/点儿；仅/有/的；已/有/的；可/再生；可/预测；可/控。

(2) 形容词性修饰词/名词性修饰词+单音节/多音节语素(通常为两个字)，一般视为两个切分单位，修饰词词表由本标准单独给出，详见附录2。例：总/计划；总/支出；低/收入；大/部分；小/部分；小/额；大/幅度；小/武器；轻/武器；最/新/计划；核/武器。注：其中，“总”，“副”作为职称时与后接词合为一个切分单位，如：总检察长。

(3) 由常用词后接修饰词时，一般看作一个切分单位。修饰词词表由本标准单独给出，详见附录2。如：中西/医学；开发/计划署；中医学；化学；物理学。

(4) 由“不”，“非”，“未”，等否定词+单音节/多音节语素，一般作为两个切分单位。例：不/明；不/满；不/可/避免；不/人道；非/正式；非/殖民化；未/经；未/成年人。

(5) 叠词，汉语以重叠变化方式构词的情况，主要有AA，AAB，ABB，AABB，A里AB，A不AB，ABAB等形式（其中A，B分别代表单字）。如果切分开不影响意思，应予以切分，本标准将单独针对叠词在表3中列出切分标准。

4.10 其他特殊片段

在汉语自然句中，会不可避免地出现非中文字符。本小节将对其进行归类，若本身非中文字符间无特殊空格划分，一般情况下连续非中文片段当作一个整体划分。

(1) 连续英文或英文+数字混合根据英文成词规则切分。即，待切分句子若有天然分隔符，则保持原有天然分隔符。例：Louis/Vuitton, Boeing/747。

(2) 中文与其他英文，罗马数字等混合，按最小语义单元划分。例：X/-/12/型/药物；β/值/增加。

(3) 其他语言文字，若无特殊空格划分，则连续非中文片段当作一个整体划分。

(4) 标点符号一般作为单独切分单位，连续相连标点，一般也要切分开，例：.../...；—/—；！/？；!/!/!。汉字或数字后若出现“.”前后，此处“.”被认作标点情况，应予以切分，否则不予以切分。例：1/.；一/.；123.45。

(5) URL和常规电话号码作为整体单独切分单元。

类别	性质	切分	示例
AA	动词	✓	走/走
	形容词	✓	甜/甜/的/点心
	名词	✓	人/人
	数量词	✓	张/张
	副词	✓	常/常
AAB	动词	×	洗洗澡
			挥挥手
ABB	形容词	×	孤单单
	数量词	✓	一/个/个
AABB	动词	×	比比划划
	形容词	×	高高兴兴
	名词	×	山山水水
	其他	×	原原本本，确确实实
A里AB	-	×	马里马虎
			糊里糊涂
			慌里慌张
A不AB	-	✓	相/不/相信
			漂/不/漂亮
			容/不/容易
ABAB	动词	✓	研究/研究
	形容词	✓	高兴/高兴
	数量词	✓	一/个/一/个
特殊形式	V一V	✓	谈/一/谈
	V了V	✓	想/了/想
	V了一V	✓	读/了/一/读

表 3: 叠词切分标准

5 总结

为了更好地适应下游信息检索与机器翻译任务，本文提出了一个相对细粒度的分词标准。仅根据分词标准进行规则模板设计难以应对汉语自动分词任务。汉语词汇的用法千变万化，随着时间的推移，信息在不断更迭，词法也会发生一定的变化，不同分词标准也各有弊益，此分词规范仅作为通用领域的总体切分准则。分词规则间互有交叉重叠，会不可避免地产生一定歧义，因此，分词标准只作为指导分词切分方向的辅助依据，黄金分割标准语料最终由实际情况和语境决定。在未来的工作中，将继续探索如何更好的在与中文相关的机器翻译任务中应用汉语分词自动，从预处理的方向出发，提升中文相关机器翻译的效果。

A 常用长词词表

本标准将提供一份补充常用长词词表，词表内的词一般情况下作为一个切分单位，为**最优先级**单独切分单位。词表由文献[1]中提供词表进行改进得来。

B 修饰词

本标准将提供一份补充修饰词词表，用作复合词切分辅助资源，具体词表如下：

修饰前缀：总、副、大、小、轻、重、最、核、性、低、高。

修饰后缀：者、性、化、家、学、法、国、人、司、科、署、局、部、员、区、组、期、团。

References

[1] Deng Cai and Hai Zhao. Neural word segmentation learning for chinese. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 409–420, Berlin, Germany, August 2016. Association for Computational Linguistics.