北京语言大学
BEIJING LANGUAGE AND CULTURE UNIVERSITY

DCC 2.0
文心语料库检索平台用户手册

创建团队：BLCU-ICALL

参与编者：朱君辉 刘鑫 师佳璐 杨麟儿 刘鹏远 杨尔弘

2023 年 4 月版

（自 2022 年 8 月起编制）

北京语言大学国家语言资源监测与研究平面媒体中心

更新记录

1.1　2023.4.21

第一部分 背景知识

第二部分 使用指南

更新记录

1.1 2023.4.21

更新了“汉语的依存句法规范”介绍。

第一部分 背景知识

一 简介

文心语料库是汉语语料库，目前包括新闻报刊语料和二语教材语料两种领域的语料。文学、微博、口语、论文等其他语料还在建设过程中。

二 主要功能特色

- 支持正则表达式；
- 支持复杂的检索表达式查询（如依存句法、命名实体、词汇难度查询等）；
- 支持按出现频次对规定的捕获内容进行统计；
- 支持查询与给定例句相同句法结构的句子；
- 支持利用元信息继续检索；
- 支持从网页上下载查询结果（txt 文件）；

.....

三 所需基础知识

（一）语料库的加工和语料库信息检索

一个语料库的功能主要与三个因素有关，一是语料库的规模，二是语料的分布，三是语料的加工程度。语料加工的深度决定这个语料库能为使用者提供什么样的语言学信息。随着语料库技术的发展，语料库的标注层次由浅入深，形成了包括分词和词性标注语料库、句法树库、语义角色标注语料库等为代表的多级加工语料库。

文心语料库中的语料已经过自动分词、自动词性标注、自动命名实体识别与自动依存句法标注。

（二）检索式

汉语研究者不仅期望语料库规模大、语料来源广，还希望语料检索软件功能强大且简单易用。在对关键词约束条件较多、检索需求较复杂的情况下，构造检索式是检索语料必不可少的一步。检索式的构造将在“第二部分 使用指南”中详细说明，在阅读之前，使用者需要熟悉以下三种背景知识：

1. **正则表达式及其常用符号：**. + * ? |

*text 匹配 text, context, pretext. (0 或多个字符）

text+ 匹配 text 和 texts （0 或 1 个字符）

text? 匹配 texts（1 个字符）

text|texts 匹配 text 或者 texts

t.*t 匹配 text, thirst, the meat 等 （任意字符）

2. CTB 的分词与词性标注规范

把语料加工技术集成在检索系统之中，是语料库检索系统的另一个特点。语料加工技术包括词语的自动切分、词性自动标注和树库自动标注等。语料加工离不开语料库规范，关于词语切分与词性标注存在几种不同的规范，文心语料库检索平台中的语料采用 The Penn Chinese TreeBank[1]（简称“CTB”）的标注规范。

[1] Xue N, Xia F, Chiou F D, Chiou F, Palmer M. The Penn Chinese TreeBank: Phrase structure annotation of a large corpus[J]. *Natural Language Engineering*, 2005, (2).

3. 汉语的依存句法规范

经过分词的语料除了标注词性以外，还可以进一步标注其他语言学属性，譬如命名实体、句法结构、短语结构等。句子的语法结构需要有形式化的方式来表达，大多数语料库采用短语结构树或依存句法树的方式，这样标注过的语料库就成为了短语树库或句法树库。标注树库是一件费时费力的工作，需要完善的标注体系和规范的标注流程以保证标注的质量，且短语结构和依存结构虽然在表现形式上不同，但是它们都是对句子语法结构的描述，在结构上存在一致性，因此很多研究人员尝试通过规则与统计的方法将短语树库转换为依存句法树库，来免去大量的人工标注工作。

对于依存句法标注规范，语言学领域有三种主要的标注方案。斯坦福依存句法标注规范（Stanford typed Dependencies，简称 SD）[2]和表层句法通用依存（Surface-syntax Universal Dependencies，简称 SUD）[3]均是基于句法规则，更适合于句法复杂性的测量；国际通用依存标注体系（Universal Dependencies，简称 UD）标注的基本原则是语义逻辑，是目前拥有语言种类最多的依存树库标注体系[4]。

目前，文心语料库检索平台中的语料经过了依存句法关系的自动标注，标注规范采用了被普遍认为最适合句法层面的 SD 标注体系。

---

[2] De Marneffe M C, Manning C D. Stanford typed dependencies manual. *Technical report*, Stanford University, 2008.

[3] Gerdes, K., Guillaume, B., Kahane, S., & Perrier, G. (2018, November). SUD or Surface-Syntactic Universal Dependencies: An annotation scheme near-isomorphic to UD. *In Proceedings of the second workshop on Universal Dependencies (UDW 2018) (pp. 66-74)*.

[4] Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajič, Christopher D. Manning, Ryan McDonald, Slav Petrov, Sampo Pyysalo, Natalia Silveira, Reut Tsarfaty, and Daniel Zeman. 2016. Universal Dependencies v1: A Multilingual Treebank Collection. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 1659–1666, Portorož, Slovenia. European Language Resources Association (ELRA).

第二部分 使用指南

一 普通检索

普通检索功能通过文心语料库检索式实现检索，支持正则表达式。常见的检索形式如[word=词&tag=词性]<advmod[...]。任意几种构成单元都可以进行组合约束。

1.1 检索表达式简介

检索式可以是字串、词串、词性等组合而成的查询模式。文心语料库检索式的构成形式分为基本项与复杂项两种，基本构成单元包括字符项、词性项、命名实体项、依存项、汉语水平等级项。基本构成单元由操作符和量词共同组成。下面依次介绍这些单元。

项目		构成形式	示例
基本项	字符项	由字符串和基本项表达式操作符构成	[word=时候]
	词性项	由词性标签和基本项表达式操作符构成	[tag=VV]
	命名实体项	由命名实体名称和基本项表达式操作符构成	[entity=PERCENT]
	依存项	由依存标签和带有方向的依存弧操作符构成	[]>nsubj[]
	汉语水平等级项	由汉语水平等级和基本项表达式操作符构成	[level=三]
复杂项		多个基本项与操作符、量词构成	/

1.1.1 操作符

操作符包括四种：

（一）基本表达式操作符：[]

（二）正则表达式操作符：. //

（三）逻辑操作符：| & = !=

（四）依存弧操作符：< > << >>

符号的含义如下：

（一） 基本表达式操作符:

基本表达式操作符用来限定基本项范围。所有的基本项都需要与基本表达式操作符[]配合使用。字符项在[word=]中进行约束，词性项在[tag=]中进行约束，命名实体项在[entity=]中进行约束，汉语水平等级项在[level=]中进行约束。具体规则将在 1.1.3~1.1.7 中进行介绍。

（二）正则表达式操作符

（1）. 通配符，代表任意单个字符

如检索式：[word=/高.兴./]

检索包含“高_兴_”的句子，“高”与“兴”后均出现一个任意字符。

（2）// 用在基本项使用正则表达式的情况中

如检索式：[tag=/N.*/]

检索包含所有以 N 开头的词性的词的句子，包括“NN”“NR”“NT”。

（三）逻辑操作符

（1）| 相当于逻辑中的“或” 关系。

如检索式：[word=只|word=仅仅]

检索包含“只”或包含“仅仅”的句子。

（2）& 相当于逻辑中的“并”关系。

如检索式：[word=只&tag=AD]

检索包含“只”且词性为副词的词的句子。

（3）= 基本项限制检索内容。

如检索式：[tag=NN]

检索包含词性为名词的词的句子。

（4）!= 基本项赋值，表示“非”。

如检索式：[word=把&tag!=BA]

检索包含“把”但并非“把”字句的句子。

（四）依存弧操作符

依存弧操作符是二元操作符，它的两边可以出现“基本项”（关于“基本项”的定义见 1.1.2）

（1）< 放置在依存标签前，示依存弧方向，入弧

如检索式：[]<nsubj[]

检索包含入弧为主语的句子

（2）> 放置在依存标签前，表示依存弧方向，出弧

如检索式：[]>dobj[]

检索包含出弧为直接宾语的句子

（3）<< 通配符，表示任何入弧，多配合量词使用

如检索式：[]<<{2,3}

检索具有任意入弧的词且入弧经过 2-3 次跳转

（4）>> 通配符，表示任何出弧，多配合量词使用

如检索式：[]>>{,2}

检索具有任意出弧的词且出弧经过 0-2 次跳转

1.1.2 量词

量词包括四种符号：+ * ? {m, n}

（1）+ 前一元素或项匹配 1 个或多个

如检索式：(>amod[])+

检索包含 1 个或多个出弧为形容词修饰语的句子。

（2）* 前一元素或项匹配 0 个或多个

如检索式：帮[]*忙

检索包含“帮”在前，“忙”在后，二者共现且中间存在 0 个或多个任意字符的句子。

（3）? 前一元素或项匹配 0 或 1 个

如检索式：帮[]?忙

检索包含“帮”在前，“忙”在后，二者共现且中间存在 0 个或 1 个任意字符的句子。

（4）{m, n} 前一项匹配 m 至 n 个元素

如检索式：帮[]{2,3}忙

检索包含“帮”在前，“忙”在后，二者共现且中间存在 2 个或 3 个任意字符的句子。

1.1.3 字符项

字符项的检索表达式有两种构成方式，检索结果一致。

（一）字符串。指不包含特殊符号和空格的连续汉字、词语。空格 SPACE 起到分隔词与词的作用（如果在词的内部使用 SPACE 是不合法的）。

如检索式：北京

如检索式：中国 北京

（二）字符串加基本项表达式操作符。指由字符串和基本项表达式操作符构成的基本项，使用[word=]。

如检索式：[word=北京]

如检索式：[word=中国] [word=北京]

1.1.4 词性项

词性可以对词进一步约束，使用“[tag=词性标签]”表示词性项，使用“[word=词&tag=词性标签]”约束检索词的词性。词性标签采用 CTB 词性体系，如下：

标签	解释	标签	解释	标签	解释	标签	解释
AD	副词	DEG	语气助词“的”	JJ	区别词	NT	时间名词
AS	助动词	DER	“得”	LB	“被”“给”等	OD	序数词
BA	“把”“将”	DEV	“地”	LC	方位词	ON	拟声词
CC	并列连词	DT	限定词	M	量词	P	介词
CD	基数词	ETC	“等”“等等”	MSP	VP 前的“所”等	PN	代词
CS	从属连词	FW	其他语言的词	NN	普通名词	PU	标点
DEC	结构助词“的”	IJ	叹词	NR	专有名词	SB	“被”“给”
VV	其他动词	VC	系动词	VE	“有”“无”“没有”	VA	表语形容词
SP	句尾小品词						

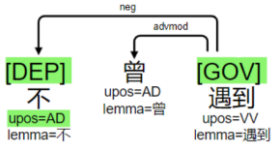
1.1.5 命名实体项

命名实体识别（Named Entity Recognition，简称 NER），又称作“专名识别”，是指识别文本中具有特定意义的实体，主要包括人名、地名、机构名、专有名词等。一般来说，命名实体识别的任务就是识别出文本中的三大类（实体类、时间类和数字类）、七小类（人名、机构名、地名、时间、日期、货币和百分比）命名实体。文心检索支持使用命名实体对检索内容进行约束。命名实体项的检索表达式写为“[entity=命名实体标签]”。命名实体，标签如下表：

标签	解释	标签	解释	标签	解释
PERSON	人名	MONEY	货币	DATE	日期
LOCATION	地名	NUMBER	数字	TIME	时间
ORGANIZATION	机构名	ORDINAL	序列号	DURATION	持续序列
MISC	杂项	PERCENT	百分比	SET	集合

1.1.6 依存项

由依存标签和带有方向的依存弧操作符构成（关于“依存操作符”的定义见 1.1.1）。文心检索系统支持带有方向的依存关系的检索，即通过描写依存弧的方向（用“<”和“>”表示）及依存关系来检索满足条件的句子，依存项的检索表达式写为“[]<依存标签[]”。如下图所示，检索类似“否定副词修饰动词”的句法结构，可以写为“[tag=AD]<advmod[tag=VV]”。



依存标签采用 UD 的依存句法标注规范。汉语中常见的依存关系见下表：

依存标签	依存关系	依存标签	依存关系	依存标签	依存关系
acl	名词修饰	compound	复合词	nsubj	名词性主语
advmod	副词修饰	conj	并列关系	nsubjpass	名词性被动主语
amod	形容词修饰	cop	系动词	nummod	数量词修饰
appos	同位语	det	限定词	parataxis	阐述关系
aux: ba	“把/将”字句	dobj	宾语	pcomp	介词补充
auxpass	助动词被动	neg	否定修饰符	pobj	介词宾语
auz	助动词	nmod	复合名词修饰	tmod	时间词修饰
case	case 标记	nn	复合名词修饰	root	根节点
ccomp	语义补充	auspass	“被”修饰		

1.1.7 汉语水平等级项

由汉语水平等级和基本项表达式操作符构成。词汇的汉语水平等级参考 2021 年 4 月由国家语委语言文字规范标准颁布并推广的《国际中文教育中文水平等级标准》（GF0025-2021）中的《词汇表》。汉语水平等级项的表达式写为“[level=希望约束的等级]”。

如检索式：[word=打&tag=VV][level=二]

检索包含动词“打”后接二级词的句子。

如检索式：[word=打&tag=VV]>dobj[tag=NN&level=二]

检索包含动词“打”的宾语为二级词且为普通名词的句子。

1.1.8 复杂项

每一基本项都可以配合不同的操作符和量词组成复杂项。具体参考“1.2 检索表达式实例”。

1.2 检索表达式实例

1.2.1 基础检索

在检索时，最简单的检索式是单个词，系统返回所有包含该检索词的语料，此时直接在检索框中输入该词或[word=该词]即可。例如：

（1）检索式：“无论如何”或“[word=无论如何]”，检索包含“无论如何”的示例；

当检索一个包含几个词的列表时，可以在词间添加符号“|”表示“或”关系，并使用辅助正则表达式的“//”。例如：

（2）检索式：“[word=/但|但是/]”，检索所有包含“但”或“但是”的实例。

同时，文心语料库也支持词性的匹配，通过输入[tag=词性]检索。例如：

（3）检索式：“[tag=/NN|NR|NT/]”或“[tag=/N.*/]”，检索所有包含所有名词（普通名词、专有名词或时间名词）的实例。

当用户希望对多个语言单位做出限制时，检索平台支持在多个语言单位之间约束顺序和间隔。当用户希望检索的语言单位直接相连时，将多个语言单位直接按照顺序构成检索式即可；当用户希望多个语言单位之间存在间隔或某些语言单位重复出现时，就需要用到正则表达式符号“*”、“+”、“?”、“{m,n}”。例如：

（4）检索式：“[word=提出][tag=NN]”，检索包含“提出+普通名词”的句子，且“提出”与普通名词在句中直接相连；

（5）检索式：“[word=只要][]+[word=就]”，检索句中先出现词“只要”，再出现“就”，并且在这两个词之间出现一个或多个词；

（6）检索式：“[tag=CD] [tag=M] [tag=/JJ|VA/]* [word=的]? [tag=/N.*/&level=三]”，检索可能带有形容词作修饰的数量短语，且作中心语的名词为《等级标准》三级词汇。

除了可以用基本项构成检索式进行简单匹配之外，还可以采用词性、词汇难度和命名实体等组合对其进行约束。例如：

（7）检索式：“[entity=ORGANIZATION&level=三]”，检索汉语水平等级为三级的机构名词。

1.2.2 依存句法检索

比如，要查询“名词性主语（nusbj 关系）+‘支持’+直接宾语（dobj 关系）”的文本，由依存句法可知，“支持”是核心动词，名词性主语和直接宾语都受其支配，但两者之间可能存在其他词语。检索表达式为“[]<nsubj[word=支持]>dobj[]”。

也可以从具体的句型出发构造检索式。例如，双宾句“给我一本书。”的依存句法结构如图 1 所示。其中，“给”是核心动词，存在“我”和“书”两个宾语。限制动词“给”必须出现的基本项表达式为“[word=给&tag=VV]”；双宾句的间接宾语大多数情况下由代词来充当，“我”作为“给”的间接宾语，通过“dobj”关系连接，依存弧方向向右，限制代词充当间接宾语的基本项写为“[tag=PN]<dobj”；“给”的直接宾语为名词“书”，通过

“dobj”关系连接，限制名词充当直接宾语的基本项写为“>dobj[tag=NN]”。将三个基本项线性地组合在一起即构成直接宾语为名词、间接宾语为代词的双宾句的检索式，检索结果如图 2 所示：

（8）检索式：“[tag=PN]<dobj[word=给&tag=VV]>dobj[tag=NN]”，检索直接宾语为名词、间接宾语为代词的双宾句。

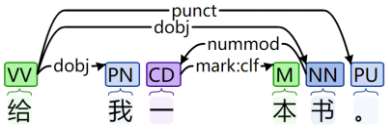


图 1　双宾句“给我一本书。”的依存句法树

图 2　依存句法检索结果示例 1

另外，弧的通配符“<<”、“>>”（表示任意依存关系）和多层弧关系“<<{m,n}”、“>>{m,n}”（表示 m 到 n 层弧）用来进行依存弧的模糊查找。以“哪怕……还……”为例，“哪怕”作为副词与句中的谓词通过“advmod”关系连接，“也”“还”同样依存于句中的谓词：

（8）检索式：[word=哪怕]<advmod[]>>[word=也|word=还]，检索包含语法结构“哪怕……也/还……”的句子，检索结果如图 3 所示。

图 3　依存句法检索结果示例 2

1.2.3 捕获

捕获是一个特殊的功能，它可以将一个语言单位捕获为变量，并支持对该变量命名，用检索式（?<name> ... ）表示，name 即表示使用者对该变量的命名。通过添加词和词性的限制条件，可以帮助用户捕获到更为精准的变量。在检索完成后，界面右侧会给出所有检索到的捕获，并按照每个变量的句子数量进行排序。在此可以对捕获进行选择来筛选检索结果，只查看包含所选捕获的句子。检索式可以包含多个捕获，也可以通过多个捕获信息来筛选检索结果。例如：

（9）检索式："（?<xx 者>[word=/.*者/]）"，检索句子中包含"者"为后缀的词，并将该词绑定在名 "xx 者"的捕获内容中高亮显示，如图 4 所示。

图 4　"xx 者"的捕获示例

值得注意的是，捕获可以与基础检索中的基本项组合使用，实现在捕获中约束依存关系、词性、词语难度等。例如：

（10）检索式：[word=刻画&tag=VV]>dobj (?<dobj> [])，检索包含动词"刻画"后接宾语的句子，并将宾语绑定在名 "dobj"的捕获内容中高亮显示，如图 5 所示。

图 5　约束依存关系的捕获示例 1

（11）检索式：[word=刻画&tag=VV]>dobj (?<dobj> [level=四])，检索包含动词“刻画”后接宾语且宾语的难度等级为四级的句子，并将宾语绑定在名“dobj”的捕获内容中高亮显示，如图 6 所示。

图 6　约束依存关系的捕获示例 2

二 模式检索

模式检索不需要用户知道底层语法表示的细节，而是通过提供一个加有简单标记的示例句子来进行查询。具有与例句的句法结构相同并符合限制条件的句子将被视为匹配，检索结果为与示例句子相匹配的句子。

3.1 模式检索表达式

模式检索的检索语言只有两种标记符号，例句中的关键词标记为捕获或锚点（用于精确匹配，检索结果中必须含有标记为锚点的词）。需要捕获的词用“( )”表示，锚点词用“$”表

示。未被标记的词，则认为是帮助句子符合语法要求、使句子完整的成分。在输入一个带标记符号的句子后，系统会返回例句的依存句法树与检索表达式呈现给用户，并依据该依存句法树匹配结果。

3.2 检索实例

例如，我们想要检索“连……也……”这样的句法结构，我们可以先为这样的句法结构造一个例句，如“他连路也走不动了”，依存句法树如下图 7 所示。

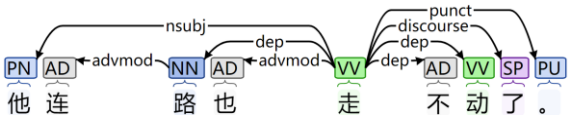


图 7 “他连路也走不动了。”的依存句法树

其中，“连”和“也”是必须要在检索结果中出现的，标记为锚点，“路”和“走”标记为捕获，即使用者想要研究的内容。那么检索式可以写为：

“他$连(路)$也(走)不动了”

与“路”和“走”在例句中发挥了同样作用的成分会被捕获，按照出现频率排列呈现在界面右侧。其返回的依存句法树、在普通检索中的检索式和检索结果如图 8 所示。

图 8 模式检索示例

四 检索结果的显示与下载

4.1 检索结果的显示单位

文心语料库的检索结果以原始语料文件（纯文本格式）中的一个自然句为单位输出显示。查询结果中，被检索项会被加粗显示。模式检索中，锚点词将被加粗显示。

4.2 检索结果的元信息

检索结果中，每条语料前的 符号代表语料的元信息，同时检索平台也支持通过语料元信息对检索结果进行二次筛选。

报刊库收录报刊的元信息包括：所在文档 ID、所在文档标题、所在文档来源、所在文档日期；

教材库收录教材的元信息包括：所在文档 ID、所在文档标题、收录日期、内容信息。

4.3 检索结果的下载

检索结果页面右上角位置有“结果下载”按钮，用户可指定下载的检索结果条数（默认为 50 条）与文件名，点击“结果下载”按钮，可将查询结果以本文文件（*.txt）格式保存至本地电脑。每句之后注明该句所在篇章名、日期等信息。

五 选择检索语料来源

文心语料库目前已入库的语料类型包括两大类：新闻报刊语料和二语教材语料。文学、微博、口语、论文等其他语料还在建设过程中。新闻报刊语料总计 16 亿字节，3950 万条句子，文本语料大小共 9.5G。主要来自《人民日报》，时间跨度为 1946 年至 2020 年。二语教材语料超 537 万字节，包含 243 万余条句子，13195 篇文章。

在检索之前，用户可在检索平台首页点击检索框下的语料来源选择对应领域的语料文本，然后输入检索式进行检索。