中文电子病历分词规范

（草案）

《中文电子病历分词规范》是根据以下资料提出的：

1.《信息处理用现代汉语分词规范》，中国国家标准 GB/T 13715-92

2.《973 当代汉语文本语料库分词、词性标注加工规范》（草案）

山西大学计算机科学系　山西大学计算机应用研究所

3.《The Segmentation Guidelines for the Penn Chinese Treebank》（3.0）

University of Pennsylvania, 2000

一、分词总则

本规范中的“分词单位”主要是词，也包括了一些结合紧密、使用稳定的词组以及在某些特殊情况下可能出现在切分序列中的孤立的语素或非语素字。

针对中文电子病历中可能出现的分词歧义，本规范结合词性进行消解，其中分词单位之间采用空格分隔，词与词性之间采用“#”分隔，词性符号的意义见第二章。分词细则包括特殊词性分词规范、组合词性分词规范及通用分词规范，各类规范的使用优先级如下表所示。

优先级	规范	适用范围
1	特殊词性分词规范	第三章中枚举的 8 类词
2	宾州树库语料(分词)	非医学领域词汇
3	组合词性分词规范	词段能够进一步被切分为两个子词的组合，并且切分后的各子词均具有词性意义
4	通用分词规范	上述规范中未提及的词类

二、词类标记集

本规范的词类标记集采用《The Part-Of-Speech Tagging Guidelines for the Penn Chinese Treebank》的大类，只针对电子病历的上层应用（例如信息抽取）增加了部分细类。

名词	NT	时间名词
	NR	专有名词
	NN	普通名词
动词	VC	系动词
	VE	“有”字动词
	VV	普通动词
形容词	VA	表语形容词
	JJ	名词修饰语
数词	CD	基数词
	OD	序数词
	DT	限定词
量词	M	/
副词	AD	/
代词	PN	/
介词	P	/
定位词	LC	/
连词	CC	并列连词
	CS	从属连词
助词	DEC	补语标记/名词化标记(de5)
	DEG	属格标记/关联标记(de5)
	DER	结果助词(de5)
	DEV	方式助词(de5)
	SP	句末助词
	AS	动态助词
	ETC	“等”字助词
	MSP	其他助词
其他词	IJ	感叹词
	ON	拟声词
	LB	长“被”结构词(bei4)
	SB	短“被”结构词(bei4)
	BA	“把”结构词（ba3）

	FW	外来词
	PU	标点符号

三、特殊词性分词规范

1. 专业术语

专业术语（包括电子病历中特有的缩写）在切分时，首先判断完整含义，如果某一子词具有明显的分类意义，则切分，例如：

血同型#JJ 半胱氨酸#NN

中枢性#JJ 面瘫#NN

否则尽量保证原有形式，不切分，例如：

自觉#VV

偏身#VV

主诉#NN

氯胺酮#NN

弥可保#NN

胰岛素#NN

上颌窦炎#NN

2. 时间名词

[1] 一周的七天，农历的初一到初十，“（大）年初一”到“（大）年初十”不切分，例如：

星期一#NT

初三#NT

[2] 年月日时分秒，按年、月、日、时、分、秒切分，例如：

1997 年#NT 3 月#NT 19 日#NT

下午#NT 2 时#NT 18 分#NT 35 秒#NT

[3] “前、后、上、下、大前、大后、头”加“天”或“上 / 下”加“月 / 周 / 星期”时，不切分，例如：

前天#NT

上周#NT

上月#NT

下星期#NT

[4] 数字与“:”或“-”结合在一起的表示具体时间的串，不切分，例如：

08:35:28#NT

2003-03-29#NT

3. 地名

表示地理区域的名称

[1] 地名后有“省、市、县、区、乡、镇、村、旗、州、都、府、道”等单字的行政区划名称时，作为一个切分单位，例如：

黑龙江省#NR

哈尔滨市#NR

青冈县#NR

[2] 地名后有表示自然区划的一个字的普通名词，“ 街，路，道，巷，里，町，庄，村，弄，堡”等，不予切分，例如：

太阳村#NR

兴安街#NR

4. 团体机构名

包括团体、机构、组织的专有名称

[1] 团体、机构、组织的专有名称其缩略式不切分，例如：

哈医大#NR

[2] 大多数团体、机构、组织的专有名称一般是短语型的，较长，且含有地名或人名等专名，对于词表中没有收录的，按词语切分开来，例如：

哈尔滨#NR 医科#NN 大学#NN

5. 量词

表示事物的单位或动词作的量。包括， 常和名词连用的名量词，有个体量词（位、辆、张），度量词（克、千米），复合量词（人次、架次、吨公里），不定量词（点、些）；以及动词量词（次、回、趟）和时量词（天、小时）等。

[1] 各类量词均要切出，例如：

3#CD 个#M 月#NN

36.2#CD　　℃#M

[2] 复合量词均不切分，例如：

次/分#M

Mmol/l #M

6. 定位词

[1] 单字定位词需要切出，例如：

左#JJ　手#NN

上#JJ　　肢#NN

内#LC　　收#VV

上#LC　　视#VV

[3] 定位词后接“侧”，“边”，“面”等后缀，整体不切分，例如：

左侧#LC

旁边#LC

注：JJ 表示定位词充当名词修饰语

7. 标点符号

独立的标点符号一律切出，例如：

蛋白#NN　:#PU　1#CD　+#PU

T#NN　:#PU　36.5#CD

8. 界限词素

界限词素附属于相邻词素以组合成词。

单字名词：校，球，院，身，科

例如：

全身#JJ

单字限定词：当

例如：

当时#NT

四、组合词性分词规范

适用于词段能够进一步被切分为两个子词的组合，并且切分后的各子词均具有词性意义。例如，“血尿”可能进一步被切分为“血”和“尿”，并且“血”和“尿”均具有名词词性。

1. ［名词+名词］

1.1 并列关系

[1] 凡是使用稳定、结合紧密的二字并列关系名词不切分，例如：

血尿#NN

口眼#NN

[2] 三字以上的结构体，其中单字部分可替换的应切分，例如：

跟#NN 膝#NN 胫#NN

肝#NN 脾#NN 肋#NN

1.2 定中关系

[1] 对 2 至 4 字组合，如其中一部分字数为 1，一般来说，整体不切分，例如：

疾病史#NN

瓣膜区#NN

脑实质#NN

腱反射#NN

肌张力#NN

高血压病#NN

病理征#NN

[2] 对两部分字数都大于或等于 2 的组合，如中间能加“的”且意义不变的切分，否则不切分，例如：

四肢#NN 肌力#NN

临床#NN 表现#NN

神经内科#NN

2. ［名词+动词］

[1] 如果动词（短语）修饰语能够插入到主谓之间，并且主语表指示，则切分，例如：

他让我很　头疼#VA

我 头#NN {很#AD} 疼#VA

[2] 主谓关系

结构体在上下文中呈整体词性时，无论字数多少，均不切分，例如：

癌变#NN

脑出血#NN

脑梗塞#NN

3. ［名词+形容词］

一般需要切分，例如：

语#NN　笨#VA

神#NN　清#VA

4. ［名词+定位词］

当满足下述所有条件时不切分，否则需要切分：

[1] 两部分均为单字

[2] 在该上下文中，名词是界限词，或者非指示词

[3] 在该上下文中，名词不被其他词修饰

例如：

国内#NN

眼前#NT

5. ［动词+名词］

5.1 动宾关系

[1] 对 2 至 4 字组合，构成动词宾式合成词时，如其中一部分字数为 1，则整体不切分，例如：

饮酒史#NN

呕吐物#NN

烧心感#NN

硬化症#NN

握痛#NN

[2] 对两部分字数都大于或等于 2 的组合，如中间能加“的”且意义不变的切分，否则不切分，例如：

止痛#NN 药物#NN

5.2 述宾关系

说明：“动词+名词”如为述宾结构的短语，名词可替换的应切分，例如：

持物#VV

进食#VV

查体#VV

构音#VV

用药#VV

伸#VV 舌#NN

咯#VV 血#NN

转#VV 头#NN

戒#VV 烟#NN

抗#VV 凝#NN

降#VV 纤#NN

6. ［动词+动词］

6.1 并列关系

双字不切分，多字切分，例如：

抬举#VV

呛咳#VV

产#VV 供#VV 销#VV

6.2 状中关系

[1] 常用的 2 字词不切分，多字词切分，例如：

伴有#VV

[2] 能愿动词与其他成分组合需要切分，例如：

可#VV 行走#VV

可#VV 有#VE

能#VV 动#VV

[3] 双字的述补结构中间插入“得”或“不”一般应切分，分别标注，例如：

走#VV 不#AD 到#VV

看#VV 得#DER 见#VV

[4] 当单字趋向动词表示抽象的趋向意义时切分，而当它们表示实在的趋向意义时不切分，例如：

步#VV 入#VV

推#VV 入#VV

掉下#VV

双字趋向动词单独切分，例如：

转#VV 进来#VV

7. ［动词+形容词］

当形容词可替换时切分，例如：

升高#VV

等#VV 大#VA

听#VV 清楚#VA

8. ［动词+介词］

动词不单独使用的不切分，其余切分， 例如：

就诊#VV 于#P

低于#VV

9. ［形容词+名词］

[1]［区别词+名词］的组合，区别词可替换的切分，例如：

副作用#NN

右#JJ 手#NN

上#JJ 肢#NN

左#JJ 下#JJ 肢#NN

大#JJ 面积#NN

[2] 多字词（4 字以上）一般需要切分，例如：

习惯性#JJ 动作#NN

周围性#JJ 面瘫#NN

[3]［形容词+“性”］的组合不切分，例如：

多发性#JJ

阵发性#JJ

10.［形容词+动词］

结合紧密、使用稳定的不切分，例如：

好转#VV

粗测#VV

多发#VV

11.［形容词+形容词］

当并列修饰名词时切分，否则不切分，例如：

左#JJ 上#JJ 肢#NN

干#JJ 湿#JJ 啰音#NN

深浅#NN 感觉#NN

12.［数词+名词］

[1] [基数词+名词] 插入量词时不改变原意，则进行切分，否则不切分，例如：

四肢#NN

序数词单独切分，例如：

二#OD 院#NN

[2] [限定词+名词] 两部分全为单字词，并且其中一部分属于界限词，则不切分，例如：

全身#JJ

半球#NN

各#DT 向#NN

[3]［基数词+词素］当词素为“来”、“余”时，不切分，例如：

头痛 30 余#CD 年

[4] 省略量词的组合，需要切分，例如：

双#CD 眼#NN

双#CD 影#NN

13.［数词+量词］

数量词组应切分为数词和量词，例如：

约#AD　3.0#CD　mm#M

血压#NN　120#CD　／#PU　70#CD　mmHg#M

多#CD 次#M

每#DT　日#M

整#DT　句#M　话#NN

14.［副词+动词］

一般需要切分，例如：

易#AD 患#VV

稍#AD 好转#VV

不#AD 能#VV

尚#AD 能#VV

才#AD 能#VV

尚#AD 可#VV

可#AD　有#VE

可#AD　见#VV

偶#AD　有#VE

多#AD　为#VC

15.［副词+形容词］

一般需要切分，例如：

较#AD 重#VA

较#AD 慢#VA

16.［代词+名词］

两部分全为单字，并且其中一部分属于界限词，则不切分，例如：

我院#NN

我科#NN

17.［代词+定位词］

两部分全为单字，则不切分，例如：

此前#NN

其中#NN

18.［否定词+形容词］

一般整体视为形容词容词不切分，例如：

不良#VA

不利#VA

不齐#VA

不详#VA

不稳#VA

不适#VA

欠清#VA

19.［多字词+前/后缀］

[1] 前缀+二字及二字以上词，若与前缀有逻辑联系的词语是与其相邻的，则不切分，例如：
肌张力#NN
腱反射#NN

[2] 二字及二字以上词＋后缀，若与后缀有逻辑联系的词语是与其相邻的，则不切分，例如：
麻木感#NN

隆隆样#JJ

心前区#JJ

药物#NN 过敏#VV 史#NN

五、通用分词规范

定义 1（组合性） 如果某词能够进一步被切分为两个子词的组合，并且切分后的各子词均具有词性意义，则该词具有组合性。

定义 2（替换性） 当某词具有组合性时，组合中的子词被其他词替换后，各部分词性均保持不变，并且新的组合仍可能出现在电子病历中，则该词具有替换性。

定义 3（还原性） 如果某词能够还原为完整语义形态，则该词具有还原性。

通用切分方案如图 1 所示，以术语“抗凝”为例，首先判断该词具有组合性和还原性，能

够预切分为“抗”和“凝”，并分别还原成动词“阻止”及“凝固”，当“凝固”替换成“发炎”时，两部分仍为动词保持不变，且新词“抗炎”同样会出现在电子病历中，所以该词具有替换性，则需要进行切分。

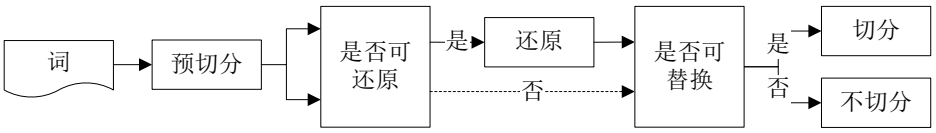

图 1　通用切分方案