《哈工大信息检索研究室同义词词林扩展版》说明

一、 英文名称

HIT IR-Lab Tongyici Cilin (Extended)

二、 词表建设

《同义词词林》的第一版和第二版的词表完全一样，收词 53,859 条。其中有很多的词已经很不常用，成为所谓的罕用词。

参照多部电子词典资源，并按照人民日报语料库中词语的出现频度，只保留频度不低于 3（小规模语料的统计结果）部分词语，可剔除 14,706 个罕用词和非常用词。经过这样的处理，《同义词词林》还剩下 39,099 个词条。为了满足自然语言处理的需要，这样规模的词典显然是少了一些，可以说远远不够。

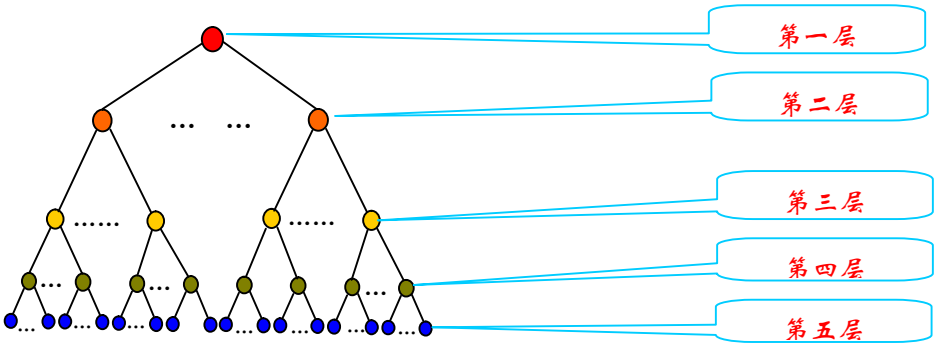
为了扩充《同义词词林》，本实验室利用很多词语相关资源，并投入了大量的人力和物力，完成了一部具有汉语大词表的《哈工大信息检索研究室同义词词林扩展版》。最终的词表包含 77,343 条词语。

二、词分类

《同义词词林》按照树状的层次结构把所有收录的词条组织到一起，，把词汇分成大、中、小三类，大类有 12 个，中类有 97 个，小类有 1,400 个。每个小类里都有很多的词，这些词有根据词义的远近和相关性分成了若干个词群（段落）。每个段落中的词语有进一步分成了若干个行，同一行的词语要么词义相同（有的词义十分接近），要么词义有很强的相关性。例如，“大豆”、“毛豆”和 “黄豆” 在同一行；“西红柿” 和 “番茄” 在同一行；“大家”、“大伙儿”、“大家伙儿” 在同一行。另外，“将官”、“校官”、“尉官” 在同一行，“雇农”、“贫农”、“下中农”、“中农”、“上中农”、“富农” 在同一行， “外商”、“官商”、“坐商”、“私商” 也在同一行，这些词不同义，但很相关。为了将词义相关的行和同义的行区分开，词典《同义词词林》在行的左端加上 “* *” 作为标记。

小类中的段落可以看作第四级的分类，段落中的行可以看作第五级的分类。这样，词典《同义词词林》就具备了 5 层结构，见图 1。随着级别的递增，词义刻画越来越细，到了第五层，每个分类里词语数量已经不大，很多只有一个词语，已经不可再分，可以称为原子词群、原子类或原子节点。不同级别的分类结果可

以为自然语言处理提供不同的服务，例如第四层的分类和第五层的分类在信息检索、文本分类、自动问答等研究领域得到应用。有研究证明，对词义进行有效扩展，或者对关键词做同义词替换可以明显改善信息检索、文本分类和自动问答系



统的性能。

词典《同义词词林》中保留下来的 39,099 条词语也保留了原有的分层结构，而新增的 36,267 条词语没有这样的结构。对于这些词，按照《同义词词林》的结构体系进行分类，工作量十分巨大。分类的某些环节可以使用机器自动完成，但是自动完成的结果不是很理想，各个环节主要还是依靠人工来完成。

三、编码

《同义词词林》只提供了三层编码，即大类用大写英文字母表示，中类用小写英文字母表示，小类用二位十进制整数表示。例如：“**Ae 07** 农民 牧民 渔民”，“**Ae 07**”是编码，“农民 牧民 渔民”是该类的标题。标题是由一个或者多个第四层的“段首（即每个段的第一个词）”组成。根据标题词可以知道小类有分成多少个第四级类，参见表 **1**。

表 1 词典结构示例

Ae07 农民 牧民 渔民
农民 农夫 农人 农 庄稼人 庄稼汉 田父 泥腿子 农家 耕夫 老乡 小农 个体农民 佃农 佃户 上中农 富裕中农 * * 菜农 棉农 茶农 烟农 蔗农 花农 药农 林农 雇农 贫农 下中农 中农 上中农 富农 自耕农 半自耕农 集体农民 人民公社社员

Ø第四级类
Ø可分为若干个第五级类

牧民 牧人 牧工

渔民 渔翁 渔家 渔夫 渔父

为了使用上的方便，对于第四级和第五级的分类也需要编码。新增的第四级和第五级的编码与原有的三级编码和并构成一个完整的编码，唯一的代表词典中的出现的词语。如：

Ba01A02= 物质 质 素

Cb02A01= 东南西北 四方

Ba01A03@ 万物

Cb06E09@ 民间

Ba01B08# 固体 液体 气体 流体 半流体

Ba01B10# 导体 半导体 超导体

编码的方法说明如下：

第四级用大写英文字母表示，第五级用二位十进制整数表示。由于第五级的分类结果需要特别说明，例如，有的行是同义词，有的行是相关词，有的行只有一个词，可以分出具体的三种情况。在使用上，有时需要对这三种情况进行区别对待，所以有必要再增加标记来分别代表着几种情形。具体的标记参见表 2。

表 2 词语编码表

编码位	1	2	3	4	5	6	7	8
符号举例	D	a	1	5	B	0	2	= \ # \ @
符号性质	大类	中类	小类		词群	原子词群		
级别	第 1 级	第 2 级	第 3 级		第 4 级	第 5 级		

表中的编码位是按照从左到右的顺序排列。第八位的标记有 3 种，分别是“=”、“#”、“@”， “=”代表“相等”、“同义”。末尾的“#”代表“不等”、“同类”，属于相关词语。末尾的“@”代表“自我封闭”、“独立”，它在词典中既没有同义词，也没有相关词。

四、词典的完善

目前推出了《哈工大信息检索研究室同义词词林扩展版》的 1.0 版本，已经可以满足很多研究领域的应用。

本实验室，还将继续组织人力对词典的功能进行必要的完善，同时修改词典分类中存在的错误。

1.0 版本秉承《同义词词林》的编撰风格，同时采用五级编码体系，提供实用的汉语大词表，以满足自然语言各个研究领域的需要。为了更好的发挥该词典的作用，本实验室拟增加更多的词语信息，如词性、读音、词频、句法关系和语义关系等。这信息的加入，将大为改观词典的结构和功能，届时也会在自然语言处理领域发挥更大的作用。