NewWordFinder 新词发现系统开发文档

自然语言处理与信息检索共享平台

Natural Language Processing & Information Retrieval Sharing Platform

http://ICTCLAS.nlpir.org/

@ICTCLAS 张华平博士

2016-2

For the latest information about NLPIR, please visit Http://ICTCLAS.nlpir.org/

访问 http://ictclas.nlpir.org/(自然语言处理与信息检索共享平台)，您可以获取 NLPIR 系统的最新版本，并欢迎您关注张华平博士的新浪微博 @ICTCLAS 张华平博士 交流。

Document Information

Document ID	NLPIR-ICTCLAS-2013-WHITEPAPER	Version	V4.0
Security level	Public 公开	Status	Creation and first draft for comment
Author	张华平	Date	Dec 19, 2013
Publisher	/	Approved by	

Version History

Note：The first version is"v0.1". Each subsequent version will add 0.1 to the exiting version. The version number should be updated only when there are significant changes, for example, changes made to reflect reviews. The first figure in the version 1.x denotes current review status by. 1. x denotes review process has passed round 1 etc .Anyone who create, review or modify the document should describe his action.

Version	Author/Reviewer	Date	Description
V1.0	Kevin Zhang	2015-2-1	first complete draft for comment. NewWordFinder

目录

NewWordFinder 新词发现系统开发文档 1
目录 4
1. NewWordFinder 新词发现系统简介 4
2. NewWordFinder 功能 C/C＋＋接口 5
2.1 NWF_Init 5
2.2 NWF_Exit 7
2.3 NWF_GetNewWords 7
2.4 NWF_GetFileNewWords 9
2.5 NWF_Result2UserDict() 10
2.6 新词发现批量处理功能 11
2.6.1 NWF_Batch_Start 11
2.6.2 NWF_Batch_AddFile 11
2.6.3 NWF_Batch_AddMem 12
2.6.4 NWF_Batch_Complete 12
2.6.5 NWF_Batch_GetResult 13
3. JNA 接口 14
3.1jna 使用分词说明 14
3.2jna 使用分词示例 14
4 运行环境 17
4.1 支持的环境 17
4.2 Linux 如何调用 NLPIR 17
5 作者简介 18

1. NewWordFinder 新词发现系统简介

新词发现系统可以自动从单篇文章、及批量文章中自动识别词典中没有出现的新词，适用于新词发现、专业词典自动生成及知识图谱中的语义新概念的自动提取。系统支持多种编码（GBK 编码、UTF8 编码、BIG5 编码）、多种操作系统（Windows，Linux， FreeBSD 等所有主流操作系统）、多种开发语言与平台（包括：C/C++/C#,Java,Python,Hadoop 等）。

我们提供各类二次开发接口，特别欢迎相关的科研人员、工程技术人员使用，并承诺非商用应用永久免费的共享策略。访问 http://ictclas.nlpir.org/(自然语言处理与信息检索共享平台)，您可以获取 NewWordFinder 系统的最新版本，并欢迎您关注张华平博士的新浪微博 @ICTCLAS 张华平博士 交流。

图 1：NLPIR/ICTCLAS 获得了钱伟长中文信息处理科学技术奖一等奖

2. NewWordFinder 功能 C/C++接口

2.1 NWF_Init

Init the analyzer and prepare necessary data for NLPIR according the configure file, same as NLPIR_Init.

```
bool NWF_Init(const char * sInitDirPath=0, int encoding=GBK_CODE,const char*sLicenceCode=0);
```

Routine	Required Header
NWF_Init	<NewWordFinder.h>

Return Value

Return true if init succeed. Otherwise return false.

Parameters

sInitDirPath: Initial Directory Path, where file Configure.xml and Data directory stored. the default value is 0, it indicates the initial directory is current working directory path

int encoding: encoding of input string, default is GBK_CODE (GBK encoding), and it can be set with UTF8_CODE (UTF8 encoding) and BIG5_CODE (BIG5 encoding).

char* sLicenceCode: license code, special use for some commercial users. Other users ignore the argument

Remarks

The **NWF_Init** function must be invoked before any operation with NLPIR. The whole system need call the function only once before starting NLPIR. When stopping the system and make no more operation, NWF_Exit should be invoked to destroy all working buffer. Any operation will fail if init do not succeed.

NWF_Init fails mainly because of two reasons: 1) Required data is incompatible or missing 2) Configure file missing or invalid parameters. Moreover, you could learn more from the log file NLPIR.log in the default directory.

Example

```
#include "NewWordFinder.h"
#include <stdio.h>
#include <string.h>

int main(int argc, char* argv[])
{
//Sample1: Sentence or paragraph lexical analysis with only one result
    if(!NWF_Init("../../",nCode))//数据在上一层目录下，默认为GBK编码的分词
    {
        printf("ICTCLAS INIT FAILED!\n");
        return ;
    }
    char sInputFile[1024]="../../test/test.TXT",sResultFile[1024];
    if (nCode==UTF8_CODE)
    {
        strcpy(sInputFile,"../../test/test-utf8.TXT");
    }

    const char *sResult=NWF_GetFileNewWords(sInputFile,50,true);//从文本文件中提取关键词

    NWF_Exit();//识别完成，系统退出，释放资源

   return 0;
}
```

Output

2.2 NWF_Exit

Exit the program and free all resources and destroy all working buffer used in NewWordFinder, same as NLPIR_Exit.

```
bool NWF_Exit();
```

Routine	Required Header
NWF_Exit	<NewWordFinder.h>

Return Value

Return true if succeed. Otherwise return false.

Parameters

none

Remarks

The **NWF_Exit** function must be invoked while stopping the system and make no more operation. And call NWF_Init function to restart NLPIR.

Example

See 5.1

Output

2.3 NWF_GetNewWords

Extract new words from paragraph.

```
NEWWORDFINDER_API const char * NWF_GetNewWords(const char *sLine,int nMaxKeyLimit=50,bool bWeightOut=false);
```

Routine	Required Header
NWF_ GetNewWords	<NewWordFinder.h>

Return Value

Return the new words list if excute succeed. otherwise return NULL.

Format as:

"科学发展观 宏观经济 " or

"科学发展观 23.80 宏观经济 12.20" with weight

Parameters

sLine, the input text.

nMaxKeyLimit, the maximum number of key words.

bWeightOut: whether the keyword weight output or not

Remarks

Example

```
#include "NewWordFinder.h"
#include <stdio.h>
#include <string.h>

int main(int argc, char* argv[])
{
//Sample1: Sentence or paragraph lexical analysis with only one result
char sSentence[2000];

const char * sResult;
if(!NWF_Init())
{
printf("Init fails\n");
return -1;
}

printf("Input sentence now('q' to quit)!\n");
scanf("%s",sSentence);

while(_stricmp(sSentence,"q")!=0)
{

const char * sKeyword= NWF_GetNewWords(sSentence);

scanf("%s",sSentence);
}
NWF_Exit();
return 0;
}
```


Output

2.4 NWF_GetFileNewWords

Extract new words from a text file.

```
NEWWORDFINDER_API const char * NWF_GetFileNewWords(const char *sTextFile,int
nMaxKeyLimit=50,bool bWeightOut=false);
```

Routine	Required Header
NWF_GetFileNewWords	<NewWordFinder.h>

Return Value

Return the keywords list if excute succeed. otherwise return NULL.

Format as:

"科学发展观 宏观经济 " or

"科学发展观 23.80 宏观经济 12.20" with weight

Parameters

sTextFile, the input text filename.

nMaxKeyLimit, the maximum number of key words.

bWeightOut: whether the keyword weight output or not

Remarks

Example

```
#include "NewWordFinder.h"
#include <stdio.h>
#include <string.h>

int main(int argc, char* argv[])
{
//Sample1: Sentence or paragraph lexical analysis with only one result
const char * sResult;
if(!NWF_Init())
{
printf("Init fails\n");
```



```
return -1;
}

const char * sKeyword= NWF_GetFileNewWords(“十八大报告.txt”);

NWF_Exit();
return 0;
}
```

2.5 NWF_Result2UserDict()

Save all extracted new words to user dictionary.
NEWWORDFINDER_API unsigned int NWF_Result2UserDict();//

Routine	Required Header
NWF_Result2UserDict	<NewWordFinder.h>

Return Value

新词识别结果转为用户词典,返回新词结果数目

Parameters

None

Remarks

Example

```
#include "NewWordFinder.h"
#include <stdio.h>
#include <string.h>

int main(int argc, char* argv[])
{
//Sample1: Sentence or paragraph lexical analysis with only one result
const char * sResult;
if(!NWF_Init())
{
printf("Init fails\n");
return -1;
}

const char * sKeyword= NWF_GetFileNewWords(“十八大报告.txt”);
```



```
NWF_Exit();
return 0;
}
```

2.6 新词发现批量处理功能

```
/*********************************************************************
*
*   以下函数为 2013 版本专门针对新词发现的过程，一般建议脱机实现，不宜在线处理
*   新词识别完成后，再自动导入到分词系统中，即可完成
*   函数以 NWF_Batch(New Word Finder Batch)开头
*********************************************************************/
```

2.6.1 NWF_Batch_Start

```
/*********************************************************************
*
*   Func Name    : NWF_Batch_Start
*
*   Description: 启动新词识别

*
*   Parameters : None
*   Returns      : bool, true:success, false:fail
*
*   Author       : Kevin Zhang
*   History      :
*                    1.create 2012/11/23
*********************************************************************/
NEWWORDFINDER_API bool NWF_Batch_Start();//New Word Indentification Start
```

2.6.2 NWF_Batch_AddFile

```
/*********************************************************************
*
*   Func Name    : NWF_Batch_AddFile
*
*   Description: 往新词识别系统中添加待识别新词的文本文件
*                    需要在运行 NWF_Batch_Start()之后，才有效
*
*   Parameters : const char *sFilename：文件名
*   Returns      : bool, true:success, false:fail
*
```



```
*   Author       : Kevin Zhang
*   History      :
*                  1.create 2012/11/23
*************************************************************************/
NEWWORDFINDER_API int    NWF_Batch_AddFile(const char *sFilename);
```

2.6.3 NWF_Batch_AddMem

```
/*************************************************************************
*
*   Func Name    : NWF_Batch_AddMem
*
*   Description: 往新词识别系统中添加一段待识别新词的内存
*                  需要在运行 NWF_Batch_Start()之后，才有效
*
*   Parameters : const char *sFilename：文件名
*   Returns      : bool, true:success, false:fail
*
*   Author       : Kevin Zhang
*   History      :
*                  1.create 2012/11/23
*************************************************************************/
NEWWORDFINDER_API bool NWF_Batch_AddMem(const char *sText);
```

2.6.4 NWF_Batch_Complete

```
/*************************************************************************
*
*   Func Name    : NWF_Batch_Complete
*
*   Description: 新词识别添加内容结束
*                  需要在运行 NWF_Batch_Start()之后，才有效
*
*   Parameters : None
*   Returns      : bool, true:success, false:fail
*
*   Author       : Kevin Zhang
*   History      :
*                  1.create 2012/11/23
*************************************************************************/
```


NEWWORDFINDER_API bool NWF_Batch_Complete();//新词

2.6.5 NWF_Batch_GetResult

```
/*********************************************************************
*
*   Func Name    : NWF_Batch_GetResult
*
*   Description: 获取新词识别的结果
*                    需要在运行 NWF_Batch_Complete()之后，才有效
*
*   Parameters : bWeightOut：是否需要输出每个新词的权重参数
*
*   Returns      : 输出格式为
*                    【新词 1】【权重 1】【新词 2】【权重 2】 ...
*
*   Author       : Kevin Zhang
*   History      :
*                    1.create 2012/11/23
*********************************************************************/
```

NEWWORDFINDER_API const char * NWF_Batch_GetResult(bool bWeightOut=false);//输出新词识别结果

Example

```
void testNewWord(int nCode)
{
    //初始化分词组件

    if(!NLPIR_Init("..",nCode))//数据在当前路径下，默认为 GBK 编码的分词
    {
        printf("ICTCLAS INIT FAILED!\n");
        return ;
    }
    char sInputFile[1024]="../test/test.TXT",sResultFile[1024];
    if (nCode==UTF8_CODE)
    {
        strcpy(sInputFile,"../test/test-utf8.TXT");
    }

    //NLPIR
```



```
    NWF_Batch_Start();//启动新词发现功能
    NWF_Batch_AddFile(sInputFile); //添加新词训练的文件，可反复添加
    NWF_Batch_Complete();//添加文件或者训练内容结束
    const char *pNewWordlist=NWF_Batch_GetResult();//输出新词识别结果
    printf("识别出的新词为：%s\n",pNewWordlist);

    strcpy(sResultFile,sInputFile);
    strcat(sResultFile,"_result1.txt");
    NLPIR_FileProcess(sInputFile,sResultFile);

    NWF_Batch_Result2UserDict();//新词识别结果导入到用户词典

    strcpy(sResultFile,sInputFile);
    strcat(sResultFile,"_result2.txt");
    NLPIR_FileProcess(sInputFile,sResultFile);

    NLPIR_Exit();
}
```

3. JNA 接口

3.1 jna 使用分词说明

Jna 编程首先根据 C 的头文件来声明对应的函数，声明后就像调用普通的 java 方法一样使用即可，详细使用例子，请见代码【注意：我们的 dll 是通用的，C、java、C#所使用的 dll 是同一个】。

3.2 jna 使用分词示例

NlpirTest 类就是对应的分词的 C 头文件的函数的声明：

```
import java.io.UnsupportedEncodingException;
import utils.SystemParas;
import com.sun.jna.Library;
import com.sun.jna.Native;

public class NlpirTest {

    // 定义接口 CLibrary，继承自 com.sun.jna.Library
    public interface CLibrary extends Library {
        // 定义并初始化接口的静态变量 这一个语句是来加载 dll 的，注意 dll 文件的路径可以是绝对路径也可以是相对路径，只需要填写 dll 的文件名，不能加后缀。
        CLibrary Instance = (CLibrary) Native.loadLibrary(
                "E://java//JNI//JnaTest_NLPIR//NLPIR", CLibrary.class);
```



```
        // 初始化函数声明
        public int NLPIR_Init(byte[] sDataPath, int encoding,
                byte[] sLicenceCode);
        //执行分词函数声明
        public String NLPIR_ParagraphProcess(String sSrc, int bPOSTagged);
        //提取关键词函数声明
        public String NLPIR_GetKeyWords(String sLine, int nMaxKeyLimit,
                boolean bWeightOut);
        //退出函数声明
        public void NLPIR_Exit();
    }
```

以下 main 函数是执行函数：

```
    public static void main(String[] args) throws Exception {
        String argu = "";
        // String system_charset = "GBK";//GBK----0
        String system_charset = "GBK";
        int charset_type = 1;
        // int charset_type = 0;
        // 调用printf打印信息
        int init_flag = CLibrary.Instance.NLPIR_Init(argu
                .getBytes(system_charset), charset_type, "0"
                .getBytes(system_charset));

        if (0 == init_flag) {
            System.err.println("初始化失败！");
            return;
        }

        String sInput = "东方网12月4日消息：2009年10月21日,辽宁省阜新市委收到举报信,举报以付玉红为首吸毒、强奸、聚众淫乱,阜新市委政法委副书记于洋等参与吸毒、强奸、聚众淫乱等。对此,阜新市委高度重视,责成阜新市公安局立即成立调查组,抽调精干力量展开调查。　　调查期间,署名举报人上官宏祥又通过尹东方(女)向阜新市公安局刑警支队提供书面举报,举报于洋等参与吸毒、强奸、聚众淫乱。11月19日,正义网发表上官宏祥接受记者专访,再次实名举报于洋等参与吸毒、强奸、聚众淫乱,引起网民广泛关注。对此辽宁省政法委、省公安厅高度重视。当日,责成有关领导专程赴阜新听取案件调查情况。为加强对案件的督办和指导,省有关部门迅速成立工作组,赴阜新督办、指导案件调查工作,并将情况上报有关部门。　　经前一段调查证明,举报事实不存在,上官宏祥行为触犯《刑法》第243条,涉嫌诬告陷害罪。根据《刑事诉讼法》有关规定,阜新市公安局已于11月27日依法立案侦查。上官宏祥已于2009年12月1日到案,12月2日阜新市海州区人大常委会已依法停止其代表资格,阜新市公安局对其进行刑事拘留,并对同案人尹东方进行监视居住。现侦查工作正在进行中。";
```



```
        String nativeBytes = null;
        try {
            nativeBytes                                                    =
CLibrary.Instance.NLPIR_ParagraphProcess(sInput, 3);
            int nCountKey = 0;
            String                      nativeByte                         =
CLibrary.Instance.NLPIR_GetKeyWords(sInput, 10,
                    false);

            System.out.print("关键词提取结果是: " + nativeByte);

            CLibrary.Instance.NLPIR_Exit();

        } catch (Exception ex) {
            // TODO Auto-generated catch block
            ex.printStackTrace();
        }

    }
```

Output:

分词结果为: 东方网/gwsh 12月/t 4日/t 消息/n : /wp 2009年/t 10月/t 21日/t ,/wd 辽宁省/ns 阜新/ns 市委/n 收到/v 举报/vn 信/n ,/wd 举报/v 以/p 付玉红/nr 为首/vi 吸毒/vi 、/wn 强奸/v 、/wn 聚众/vn 淫乱/vn ,/wd 阜新市/ns 委/ng 政法委/n 副/b 书记/n 于洋/nr 等/udeng 参与/v 吸毒/vi 、/wn 强奸/v 、/wn 聚众/vd 淫乱/v 等/udeng 。/wj 对/p 此/rzs ,/wd 阜新/ns 市委/n 高度/d 重视/v ,/wd 责成/v 阜新市/ns 公安局/n 立即/d 成立/vi 调查组/n ,/wd 抽调/v 精干/a 力量/n 展开/v 调查/vn 。/wj 调查/v 期间/f ,/wd 署名/vi 举报人/n 上官/nr1 宏祥/nr2 又/d 通过/p 尹东方/nr (/wkz 女/b)/wky 向/p 阜新市/ns 公安局/n 刑警/n 支队/n 提供/v 书面/b 举报/vn ,/wd 举报/v 于/p 洋/ag 等/udeng 参与/v 吸毒/vi 、/wn 强奸/v 、/wn 聚众/vn 淫乱/vn 。/wj 11月/t 19日/t ,/wd 正义/n 网/n 发表/v 上官/nr1 宏祥/nr2 接受/v 记者/n 专访/vn ,/wd 再次/d 实/a 名/q 举报/v 于/p 洋/ag 等/udeng 参与/v 吸毒/vi 、/wn 强奸/v 、/wn 聚众/vn 淫乱/vn ,/wd 引起/v 网民/n_new 广泛/ad 关注/v 。/wj 对/p 此/rzs 辽宁省/ns 政法委/n 、/wn 省/n 公安厅/n 高度/d 重视/v 。/wj 当日/t ,/wd 责成/v 有关/vn 领导/n 专程/d 赴/v 阜新/ns 听取/v 案件/n 调查/vn 情况/n 。/wj 为/p 加强/v 对/p 案件/n 的/ude1 督办/vn 和/cc 指导/vn ,/wd 省/n 有关/vn 部门/n 迅速/ad 成立/vi 工作组/n ,/wd 赴/v 阜新/ns 督办/v 、/wn 指导/vn 案件/n 调查/vn 工作/vn ,/wd 并/cc 将/p 情况/n 上报/vi 有关/vn 部门/n 。/wj 经/p 前/f 一/m 段/q 调查/v 证明/v ,/wd 举报/v 事实/n 不/d 存在/v ,/wd 上官/nr1 宏祥/nr2 行为/n 触犯/v 《/wkz 刑法/n 》/wky 第243/m 条/q ,/wd 涉嫌/v 诬告/v 陷害/v 罪/n 。/wj 根据/p 《/wkz 刑事诉讼法/n 》/wky 有关/vn 规定/n ,/wd 阜新市/ns 公安局/n 已/d 于/p 11月/t 27日/t 依法/d 立案/vi 侦查/v 。/wj 上官/nr1 宏祥/nr2 已

/d 于/p 2009年/t 12月/t 1日/t 到/v 案/ng ,/wd 12月/t 2日/t 阜新市/ns 海州区/ns 人大/n 常委会/n 已/d 依法/d 停止/v 其/rz 代表/n 资格/n ,/wd 阜新市/ns 公安局/n 对/p 其/rz 进行/vx 刑事/b 拘留/vn ,/wd 并/cc 对/p 同/p 案/ng 人/n 尹东方/nr 进行/vx 监视/vn 居住/vn 。/wj 现/tg 侦查/v 工作/vn 正在/d 进行/vx 中/f 。/wj

关键词提取结果是：阜新市公安局#有关部门#案件调查#赴阜新#阜新市委#调查#举报#有关#进行#尹东方#阜新#

4 运行环境

4.1 支持的环境

1. 可以支持 Windows、Linux、FreeBSD 等多种环境，支持普通 PC 机器即可运行。

2. 支持 GBK/UTF-8/BIG5

4.2 Linux 如何调用 NLPIR

1）与 window 下一样编程；

2）Makefile 的命令如下：

test: ../../../Src/ICTCLAS2013/example-c/Example-C.cpp ../../../Src/ICTCLAS2013/include/NewWordFinder.h

g++ ../../../Src/ICTCLAS2013/example-c/Example-C.cpp -L. -lpthread -L../../../bin/ICTCLAS2013 -lNLPIR -Wall -Wunused -O3 -DOS_LINUX -o ../../../bin/ICTCLAS2013/example

其中 Example-C.cpp 是测试使用 NLPIR 的程序；因为 NLPIR 进行了多线程的安全保护设计，需要调用多线程的库，即-L. –lpthread。调用 nlpir 的部分为：-L../../../bin/ICTCLAS2013 –lNLPIR 第一部分为路径，后面为 libNLPIR.so 对应的名称-lNLPIR。具体可以参见

5 作者简介

张华平 博士 副教授 研究生导师
大数据搜索挖掘实验室 主任
地址：北京海淀区中关村南大街 5 号 100081
电话：+86-10-68918642 13681251543(助手电话)
Email:kevinzhang@bit.edu.cn
MSN: pipy_zhang@msn.com;
网站: http://www.nlpir.org (自然语言处理与信息检索共享平台)
http://www.bigdataBBS.com (大数据论坛)
博客:http://hi.baidu.com/drkevinzhang/
微博:http://www.weibo.com/drkevinzhang/

Dr. Kevin Zhang (张华平, Zhang Hua-Ping)
Associate Professor, Graduate Supervisor
Director, Big Data Search and Mining Lab.
Beijing Institute of Technology
Add: No.5, South St.,Zhongguancun,Haidian District,Beijing,P.R.C
PC:100081
Tel: +86-10-68918642 13681251543(Assitant)
Email:kevinzhang@bit.edu.cn
MSN: pipy_zhang@msn.com;
Website: http://www.nlpir.org (Natural Language Processing and Information Retrieval Sharing Platform)
http://www.bigdataBBS.com (Big Data Forum)
Blog:http://hi.baidu.com/drkevinzhang/
Twitter: http://www.weibo.com/drkevinzhang/