NLPIR-Parser: Making Chinese and English Semantic Analysis Easier and Complete

Huaping Zhang[1], Jun Miao[2], Ziyu Liu[1], Ian Logan Wesson[1], Jianyun Shang[1]

[1]Beijing Institute of Technology, Beijing – kevinzhang@bit.edu.cn, liuziyu2017@nlpir.org, GeometricAlgorithm@gmail.com, shangjia@bit.edu.cn

[2]Sun Yat-sen University, Guangzhou (Zhuhai campus) – miaoj5@mail.sysu.edu.cn

Abstract

This paper presents NLPIR-Parser, which is an enhanced semantic analysis platform and a total solution for mining Chinese and English corpora. By providing a powerful search engine with information filtering, NLPIR-Parser can help the user web-crawl all possible online data sources of topic keywords and website URLs; the NLPIR-Parser includes tokenization and Part-Of-Speech (POS) tagging, named entity recognition and new words identification for natural language processing; it also includes text classification and clustering. The NLPIR-Parser can easily and effectively apply the whole semantic analysis process for huge corpora without any further programming and any private data leaking; more importantly, the platform is integrated with the Chinese word segmentation system known as ICTCLAS; last but not least, it can present uniform tokenization and POS tagging for any mixture of English and Chinese texts. Taking "the Belt and Road Initiative" as exemplification, this paper will illustrate some functions in the NLPIR-Parser. Based on semantic analysis results, some social analysis conclusion will be achieved.

Keywords: Chinese word segmentation; tokenization; new words identification; semantic Analysis.

Résumé

Cet article présente la plate-forme d'analyse sémantique NLPIR-Parser qui permet l'étude complète de corpus chinois et anglais. Grâce à son puissant moteur de recherche et à un filtrage d'information, NLPIR-Parser sert l'utilisateur dès le prétraitement des donnés, en recherchant toutes les sources en ligne, à l'aide de mots-clés ou par URL de sites Web. Pour le traitement du langage naturel, NLPIR-Parser inclut la lemmatisation et l'étiquetage de la partie du discours (POS), il reconnaît les noms des entités et peut identifier de nouveaux mots. Durant l'analyse textuelle, il classifie et regroupe les textes. Son processus d'analyse sémantique fonctionne facilement et efficacement, sur de grands corpus, sans qu'il y ait besoin de recourir à d'autre programmation, ce qui améliore d'autant la sécurité des données. Plus important encore, la plate-forme a intégré le système de segmentation du chinois en ICTCLAS ; enfin et surtout, NLPIR-Parser peut segmenter et étiqueter des textes qui mélangent de textes anglais et chinois. En prenant « l'initiative de la ceinture et de la route » (*la nouvelle route de la soie*) comme exemple, nous illustrerons certaines fonctions de la plateforme et montrerons comment les résultats de l'analyse sémantique reflètent une situation sociale.

Mots clés : segmentation du chinois en mot; lemmatisation; identification de nouveaux mots; analyse sémantique.

1. Introduction[1]

[1] This work was supported by National Social Science Foundation of China (Grant No. 17XYY012), National Science Foundation of China (Grant No. 61772075), and Ministry of Science and Technology of China (Grand No 2018YFC0832304). The authors would like to thank Kim Gerdes, Sylvie Royer for their constructive criticism of this paper.

Semantic analysis on corpora of a wide range of domains and genres is becoming more and more important with the development of social media. Although there are some popular corpus analysis tools, obtaining a complete and precise one is still a great bottleneck for common users.

While processing a corpus, there is generally a large requirement for manual annotation, which is in urgent need of fast and simple natural language processing tools. It is also often difficult to make a uniform text mining and comparison for different languages, especially for a pair of Western and Eastern languages, such as Chinese and English. This paper presents our solutions, and our analysis tool-platform: NLPIR-Parser[2] (Zhang and Shang, 2019). After a brief description of its architecture, we will take "the Belt and Road Initiative" in Chinese and in English as sample topic, and address the differences of NLPIR-Parser and other natural language processing (NLP) related tools.

2. Current challenges of NLP practical application

Natural language processing (NLP) technologies have grown rapidly in the past 30 years, making great progress. But, for ordinary researchers, there are still the following challenges:

2.1. A high learning curve

Technicians are usually required to participate in the development of NLP tools, and the cost of learning is (often insurmountably) high for the majority of targeted end-users, which are researchers with backgrounds in liberal arts. For corpora processing projects, the cost of this is high and time-consuming, and the post-processing of the tagged content, such as clustering, classification, and visualization, cannot be completed manually. Computer software is needed to deal with this. Ready-made commercial or open source tools can be used, but their effectiveness is limited because of the high learning-curve of using these tools, especially for liberal arts personnel, who are often in need of these tools, especially in the case of word processing.

2.2. Over-Specialization

Most of the existing tools have limited functionality, and lack the versatility of a full-chain semantic analysis tool. To better understand the fragmented sets of functionality, consider the following array of NLP-related tools: **urllib2**, **Scrapy** and **Pyspider** provide information crawling tools; **WordSmith** (Scott, 2008), **AntConc** (Anthony, 2005), and **Lexico** (Salem, 1990/2018) provide statistical and analytical functions ; **jieba**[3] provides Chinese word segmentation tools; and, **LTP** - the language technology platform of Harbin Institute of Technology[4] - provides tools such as Chinese word segmentation, part of speech tagging, named entity recognition, dependency parsing, semantic corner color tagging and so on. However it is necessary to construct HTTP requests according to API specifications to obtain the analysis results online.

To solve the above practical challenges that many researchers encounter, and to satisfy the urgent need for a comprehensive client-side NLP tool, the NLPIR-team has developed a full-chain NLP platform: NLPIR-Parser[5] (Zhang and Shang, 2019).

[2] It can be accessed in NLPIR.org or Github https://github.com/NLPIR-team/NLPIR/tree/master/NLPIR-Parser
[3] https://github.com/fxsjy/jieba
[4] https://www.ltp-cloud.com/
[5] NLPIR online demo is : http://ictclas.nlpir.org/nlpir/

2.2. Ambiguity between different Language Standards

Each language has a different word and Part-Of-Speech (or POS) standard, therefore it is hard to make a uniform text mining and comparison for different languages, especially for a pair of Western and Eastern languages, such as Chinese and English. For instance, an English verb can be divided into different POS, like “VB” (Verb, *base form*), “VBD” (Verb, *past tense*), “VBG” (Verb, *gerund or present participle*), “VBN” (Verb, *past participle*) (Upenn, 2003). However, the verb in Chinese has only one form “v” in Peking University standard (Yu *et.al.*, 2002).

2.3. The threat of data leakage

There is usually a threat of data leakage in the corpus knowledge resources to be processed, because, at present, most natural language processing organizations provide a natural language processing cloud service platform, such as in the cases of NLP cloud service from Google or Baidu, which require the users to upload the corpus to be processed. In many cases, most of the researchers' corpora are collected and annotated with a large cost of manpower and resources, sometimes even being the accumulation of lifelong painstaking efforts. This makes the corpus much more expensive and time-consuming to process. After the data resources stored in the cloud are separated from the uploaders, there is no legal protection for the right of the data. This creates a huge hidden danger of data leakage and abuse, which makes many users shy away.

3. NLPIR-Parser Architecture

The NLPIR-Parser architecture is illustrated in *Figure 1*. It comprehensively integrates data import tools, NLP tools, information extraction tools, and text mining tools. During the data import stage, we can use topic words, website URLs or any document format with any encoding to get the data.

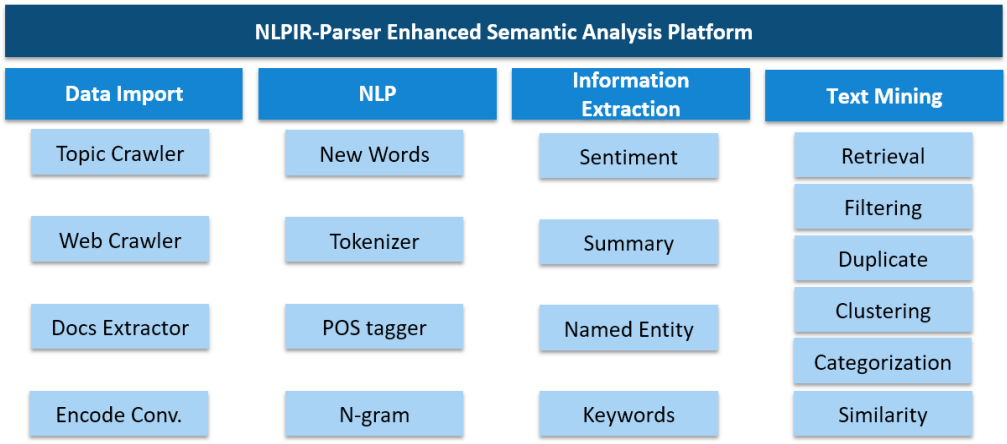

Figure 1. NLPIR-Parser Architecture

At the NLP stage, *new words detection*, *tokenizer* (Chinese word segmentation and English tokenization), *POS tagging*, and *n-gram* language modeling can be automatically done. At information extraction stage, NLPIR-Parser can extract *sentimental words*, *keywords*, *named entities* and perform *text summarization*. In the *text mining* stage, NLPIR-Parser can also do *information retrieval* and *filtering*, *document reduplication detection*, *text clustering* and *categorization*, and *text similarity computation*. So, from the data stage to the deep analysis,

the NLPIR-Parser provides a full-chain semantic analysis platform, without any connection to an outside server or cloud. It offers a safe and reliable tool to its users.

We have an online demo version (http://ictclas.nlpir.org/nlpir/), and the client can use it directly, but there is also a client-side *application programming interface* (API) for further development. Figure 2 is a screenshot of the client-side platform. It is easy to use for processing a large amount of language and media information, even for liberal arts students who may have no technical background. Additionally, the NLPIR-Parser API can be seamlessly integrated into all kinds of complex application systems for end users. This API is also compatible with almost all popular operating systems, such as Windows, Linux, Android, Maemo5, FreeBSD, and more. It can also be invoked by various development languages such as Python, Java, C, C# and PHP.

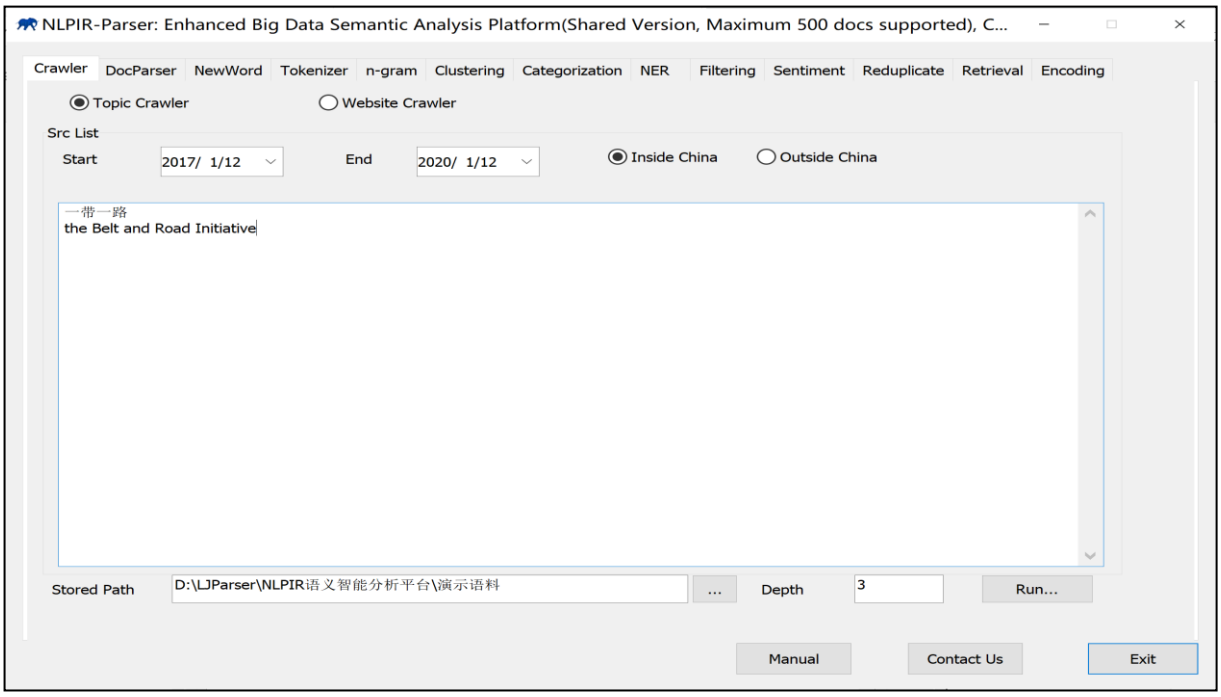

Figure 2. NLPIR-Parser Client Program

4. Illustration Using "the Belt and Road Initiative" as Input

In the following section, we will illustrate the details of the NLPIR-Parser functionality with "the Belt and Road Initiative" in English and "一带一路" in Chinese as an example input. The "Belt and Road Initiative" is a global development strategy proposed by the Chinese government in 2013 with the aim of improving regional integration, increasing trade and stimulating economic growth.

4.1. Data Import

We have 3 ways to import data sources:

Topic Crawling: in the crawler module, the user can click the radiobox labeled "Topic Crawler", and input each keyword line by line. Here, the user can input two lines: "一带一

路" and "the Belt and Road Initiative", as seen in figure 2 above. The system will submit the keywords as a query to different search engines, such as Baidu, Microsoft Bing and Google. Then the crawler will automatically extract information from all returned search result pages.

Website Crawling: in the crawling module, the user can click the radiobox labeled "Website Crawler", and input all website URLs of interest, one by one. NLPIR will crawl all webpages inside the website list. The crawling depth is, by default, set to 3, and the crawling speed is set, by default, to achieve 50 pages/second.

Documents Import: in the "DocParser" module, the user can upload documents with many possible formats, such as doc, docx, pdf, xls, ppt, html and text. If the input is plain text, NLPIR parser can convert it to utf-8 format by using "Encoding" modules.

To demonstrate NLPIR-Parser's semantic analysis, 4 Chinese documents with their English translations (all listed below) being crawled in parallel will be used as input

-Chinese and English versions of the keynote speech by Xi Jinping, President of the People's Republic of China, at the Opening Ceremony of the First Belt and Road Forum for International Cooperation on May 14, 2017. (Xi, 2017)

-Chinese and English versions of accomplishment results during the First Belt and Road Forum for International Cooperation.

-Chinese and English version of the keynote speech by Xi Jinping, President of the People's Republic of China, at the Opening Ceremony of the Second Belt and Road Forum for International Cooperation on April 26, 2019. (Xi, 2019)

-Chinese and English versions of the accomplishment results in the Second Belt and Road Forum for International Cooperation Forum.

4.2. Natural Language Processing

NLPIR-Parser provides new words identification, tokenization, POS tagging and n-gram language modeling when performing NLP.

4.2.1. New Words Identification

In the open domain or on social media, there exists a huge number of possible words. These new words constitute an important part of representing language features. The NLPIR-Parser uses a scalable and parallel approach to new word extraction. It uses a conditional random field (CRF) model to segment the corpus to extract the new word candidates, then uses a bigram language model to segment the corpus again in order to extract the features of mutual information and adjacent entropy. This approach has a linear time complexity and reduces memory usage as well as improves the performance of parallelization by reducing the dependency on a global state. The experiments proved that this algorithm is both effective and accurate compared to traditional approaches and is well-suited for processing large corpora in the web. (Zhang and Shang, 2017)

In the NewWord module, the user can browse the corpus path and automatically detect new words in both Chinese and English. In this example, 113 new words are detected from 4 Chinese documents and 4 English documents. The top 10 Chinese new words and 3 English new words (or phrases) are given in *Table 1*. Here, the "Word" column refers to the newly detected words, and the "Translation" column is the literal translation into Chinese. The "Frequency" column is the total number of occurrences of the word in the corpus, and the "Weight" column is computed using adjacent entropy. The result is sorted by weight. From

the table, we find that the new words can effectively represent linguistic features of the “the Belt and Road Initiative”.

Table 1. New words identified in the corpus (mixed with Chinese and English)

Word	Translation	Weight	Frequency
一带一路	*One Belt and One Road*	37.9	102
南南合作	*South-South cooperation*	27.3	8
基础设施	*infrastructure*	25.68	19
合作规划	*cooperative planning*	22.23	12
联合国教科文组织	*UNESCO*	20.8	7
沿线国家	*Countries along the routine*	20.79	13
silk road		20.67	29
知识产权	*intellectual property*	20.62	16
互联互通	*Interconnection*	20.51	11
高峰论坛	*summit forum*	20.13	22
智库	*think tank*	19.94	16
maritime silk road		15.36	5
mutual learning		10.37	3

4.2.2 Tokenization and POS tagging

The NLPIR-Parser integrates ICTCLAS for tokenization (Zhang *et al.*, 2003), which supports both Chinese word segmentation and English tokenization. ICTCLAS ranked top in the SIGHAN first international bakeoff in 2003 and was awarded with the first prize award given by the Chinese Information Processing Society.

```
齐心/ad 开创/v 共/d 建/v “/wyz 一带一路/n_new ”/wyy 美好/a 未来/t
（/wkz 2019年4月26日/t，/wd 北京/ns ）/wky
尊敬/v 的/ude1 各位/rr 国家/n 元首/n，/wd 政府/n 首脑/n，/wd
各位/rr 高级/a 代表/n，/wd
各位/rr 国际/n 组织/v 负责人/n，/wd
女士/n 们/k，/wd 先生/n 们/k，/wd 朋友/n 们/k：/wm
上午/t 好/a！/wt “/wyz 春秋/n 多/m 佳/a 日/ng，/wd 登高/vi 赋/ng 新诗/n 。/wj ”/wyy 在/p 这个/rz 春意盎
然/vl 的/ude1 美好/a 时节/n，/wd 我/rr 很/d 高兴/a 同/p 各位/rr 嘉宾/n 一道/d，/wd 共同/d 出席/v 第二/m 届/q
“/wyz 一带一路/n_new ”/wyy 国际/n 合作/vn 高峰论坛/n_new 。/wj 首先/c，/wd 我/rr 谨/d 代表/v 中国政府/nt
和/cc 中国/ns 人民/n，/wd 并/cc 以/p 我/rr 个人/n 的/ude1 名义/n，/wd 对/p 各位/rr 来宾/n 表示/v 热烈/a
的/ude1 欢迎/vn！/wt
两/m 年/qt 前/f，/wd 我们/rr 在/p 这里/rzs 举行/v 首/m 届/q 高峰论坛/n_new，/wd 规划/v 政策沟通/n_new 、/wn
设施联通/n_new 、/wn 贸易/vn 畅通/an 、/wn 资金融通/n_new 、/wn 民心相通/n_new 的/ude1 合作/vn 蓝图/n
。/wj 今天/t，/wd 来自/v 世界/n 各地/rzs 的/ude1 朋友/n 再次/d 聚首/vi 。/wj 我/rr 期待/v 着/uzhe 同/p 大家/rr 一
起/s，/wd 登高望远/vl，/wd 携手/vd 前行/v，/wd 共同/d 开创/v 共/d 建/v “/wyz 一带一路/n_new ”/wyy
的/ude1 美好/a 未来/t 。/wj
```

Figure 3. Chinese word segmentation result on President Xi’s Keynote Speech (in Chinese) in 2019

In the tokenizer modules, ICTCLAS can adapt to a new domain if the user imports the new words explained in section 4.2.1 (the new words file can be modified with POS labels as required). It is important to point out that ICTCLAS supports both Chinese and English, and

that both languages use the same system and POS tagging set (ICTPOS 3.0[6], ICTCLAS, 2001). *Figure 3* shows the Chinese word segmentation and POS tagging result. The precision can achieved can be over 98% on an open domain after importing new words (Zhang and Shang, 2017).

In the following, *Figure 4* shows the result of English tokenization and POS tagging.

```
Working/vi Together/d to/pba Deliver/v a/rzv Brighter/a Future/n For/p Belt/n and/c Road/initiative
Cooperation/documents 18/m :/wm 24/m ,/wd April/t 26/m ,/wd 2019/m Keynote/n Speech/n by/p H.E/n ./wj
Beijing/a ,/wd 26/m April/t 2019/m Your/rr Excellencies/n Heads/n of/p State/n and/c Government/n ,/wd
Your/rr Excellencies/n High/a -/wp level/n Representatives/n ,/wd Your/rr Excellencies/n Heads/n of/p
International/a Organizations/n ,/wd Ladies/n and/c Gentlemen/n ,/wd Friends/n ,/wd Good/a morning/n !/wt
As/p a/rzv line/n of/p a/rzv classical/a Chinese/side poem/n goes/v ,/wd Spring/n and/c autumn/n are/vshi
lovely/a seasons/n in/p which/rzs friends/n get/v together/d to/pba climb/n up/pbei mountains/n and/c
write/v poems/n ./wj On/p this/r beautiful/a spring/n day/n ,/wd it/rzt gives/v me/rzv great/a pleasure/n
to/pba have/vyou you/rzt with/p us/rzv here/d at/p the/rzt Second/m Belt/n and/c Road/initiative Forum/n
for/p International/a Cooperation/documents BRF/n )/wky ./wj On/p behalf/n of/p the/rzt Chinese/side
government/n and/c people/n and/c in/p my/rr own/rzs name/n ,/wd I/rzt extend/v a/rzv very/cc warm/a
welcome/a to/pba you/rzt all/a !/wt
```

Figure 4. English tokenization and POS result on President Xi's Keynote speech（in English） in 2019

In this way, the NLPIR-Parser can be used for mixed Chinese and English corpora.

4.2.3. N-Gram language modeling

In the "n-gram" module, the NLPIR-Parser can generate unigram and bigram language modeling results.

Figure 5 shows that the corpus includes 2329 unique open class words, such as nouns, verbs, adjectives and numbers. The average frequency is 3.86. The columns are word, POS, frequency, unigram probability, entropy value and its translation automatically generated by the system, respectively. Unigram results can help the users quickly understand the word distribution.

As we can see in *Figure 5*, China (中国) is the country to propose the "one Belt and one Road Initiative" (一带一路). It is a national project of cooperation (合作) with other countries in the international framework (国际). The main objective is to jointly (共同) promote the economic development (经济发展). The Memorandum of understanding (谅解备忘录) between countries is signed. To develop (发展) the projects, the banks (银行) will invest and provide financial support (投资).

[6] POS tagging rules: http://ictclas.nlpir.org/nlpir/html/readme.htm

Word Count: 2329, Average Frequency: 3.863461					
Word	POS	Freqency	Unigram	Entropy	Translation
中国	ns	228	0.025339	0.093131	China; Chinese (adj.)
一带一路	nz	203	0.022561	0.08554	
合作	vn	195	0.021671	0.08304	cooperate; collaborate; work together 技
签署	v	155	0.017226	0.069961	sign; affix; subscribe ~联合公报 sign a jo
国家	n	114	0.012669	0.055347	country; state; nation 发展中~ developi
国际	n	110	0.012225	0.053842	international ~地位 international status
发展	vn	88	0.00978	0.045256	①（变化）develop; expand; grow; deve
项目	n	74	0.008224	0.039481	①（门类） item; project 重复~ duplicat
建设	vn	70	0.00778	0.037779	build; construct; construction (n.) 社会主
政府	n	62	0.00689	0.034298	government 廉洁高效的~ a clean and e
共同	d	57	0.006335	0.032065	①（属于大家的） common ~关心的问
谅解备忘录	nz	49	0.005446	0.028388	
要	v	48	0.005335	0.027919	①（重要） important; essential ~事 an
经济	n	48	0.005335	0.027919	①〈经〉 economy 国民~ national econ
银行	n	47	0.005223	0.027447	bank 中国建设~ China Construction Ban
发展	v	46	0.005112	0.026973	①（变化）develop; expand; grow; deve
投资	vn	41	0.004557	0.024565	①（投入资金） invest; put money into

Figure 5. Screenshot of an excerpt of unigram Result in Excel

Figure 6 shows that the corpus includes 7811 bigram pairs. The columns are previous word, next word, co-occurrence frequency, bigram probability, and bigram entropy value, respectively. Bigram results can help the users find frequent pairs, which are useful for segment analysis. The results are: 中国国家 *of Chinese government /of China*, 签署关于 *to signed agreements on*，建一路一带 *to carry out the Belt and Road Initiative.*

Bigram pair:7811				
Previous Word	Next Word	Co-occurrence	Bigram Prob.	Bigram Entropy
中国	国家	48	0.210526	0.027919
签署	关于	48	0.309677	0.027919
建	一带一路	38	0.926829	0.023089
一带一路	建设	38	0.187192	0.023089

Figure 6. Screenshot of an excerpt of bigram Result in Excel

4.3. Information Extraction

The information extraction toolset includes: keywords extraction, named entities extraction, text summarization and sentimental analysis. Automatic summarization can automatically extract the essence of the content of single or multiple articles, which is convenient for users who need to quickly browse the text content. Entity extraction can automatically extract content abstracts, named entities including person names, place names, organization names, time, and subject keywords from single or multiple articles; it is convenient for users who need to quickly browse text content.

With input text “意大利本月正式加入‘一带一路’，美国反应非常搞笑” (*Italy officially announced to join in ‘the Belt and Road Initiative’ this month, the reaction of USA is funny*) as an example, it can seen that the extracted information, particularly for the *named entities* (NER), is very precise. NLPIR-Parser’s NER module uses a role based model (Zhang et *al*, 2003).

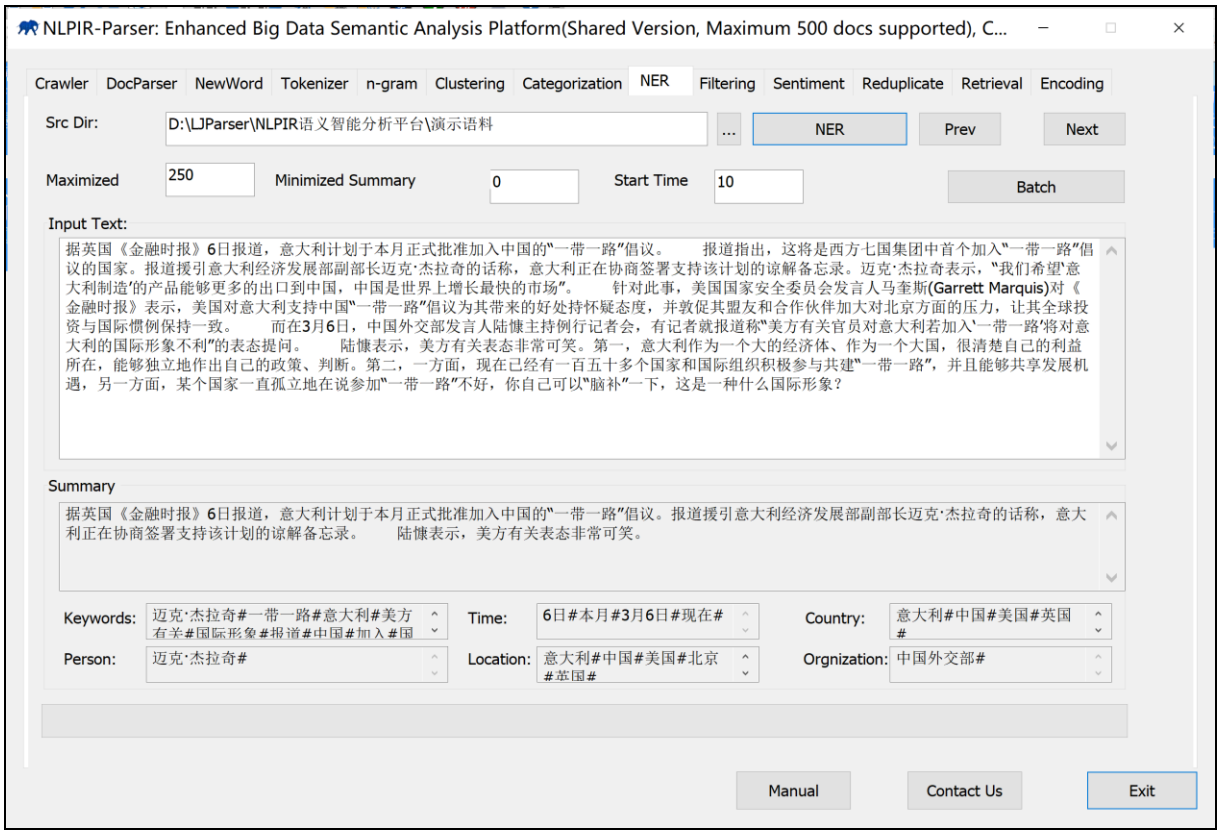

Figure 7. Information Extraction Sample

As seen in *Figure 7*, the extracted result in the named entities (NER) tab has the following output:

Summary : 据英国《金融时报》6 日报道，意大利计划于本月正式批准加入中国的”一带一路”倡议。报道援引意大利经济发展部副部长迈克·杰拉奇的话称，意大利正在协商签署支持该计划的谅解备忘录。陆慷表示，美方有关表态非常可笑。
(*According to 6th report in Financial Times in England, Italy planed to approve officially join the “the Belt and Road Initiative” at the end of this month. The news cited the speech of Michele Geraci, vice minister in Italian Economic Ministry. Italy was negotiating to sign the initiative MOU. Lu Kang expressed that US declaration was ridiculous.*)

Keywords : 迈克·杰拉奇(Michele Geraci)#一带一路(the Belt and Road Initiative)#意大利(Italy)#美方有关(related to US)#国际形象(international fame)#
Time : 6 日(6th)#本月(this month)#3 月 6 日(March 6)#现在(Now)#
Country : 意大利(Italy)#中国(China)#美国(US)#英国(England)#
Person： 迈克·杰拉奇(Michele Geraci)
Location： 意大利(Italy)#中国(China)#美国(US)#北京(Beijing)#英国(England)#
Organization： 中国外交部(Chinese Ministry of Foreign Affairs)

The extracted information includes the time, the country (*Italy*, *China*, *US* and *England*), the person (*Michele Geraci*), the location, and also the organization. Automatic summarization approaches can be found in (Li et *al*, 2014). Extracted keywords includes not only known words in the lexicon, but also out-of-vocabulary new words (Zhang and Shang, 2019).

NLPIR can also carry out the sentimental analysis. It can be performed when given a whole document or set of entities, such as people, organizations or brand names. During sentimental analysis, for the pre-designated analysis object, the system automatically analyses the emotional tendency of a large number of documents. It also calculates emotional polarity and emotional value measurements, and gives positive and negative scores and sentence examples from the original text. NLPIR emotional analysis can distinguish between the rich set of emotional categories, including not only positive and negative aspects, but also the specific emotional attributes, such as: good, joy, surprise, anger, malice, sadness and fear. It can also provide emotional analysis of any given set of entities.

4.4. Text Mining

Text mining includes text classification, clustering, redepulication detection, information retrieval, and filtering.

4.4.1. Text classification

The text classification problem is to classify a document into one or more of several predefined categories. Automatic text classification uses computer programs to achieve such classification. Text classification can automatically identify and train classifications from massive documents according to pre-specified rules and samples.

NLPIR in-depth text classification can be used in news classification, resume classification, email classification, office document classification, and regional classification, just to name a few. Additionally, it can perform text filtering by quickly identifying and filtering information that meets special requirements from a large number of texts. This functionality can also be used in brand report monitoring, spam shielding, sensitive information censorship, and many other applications of text-classification.

NLPIR uses deep neural networks to train the classification system comprehensively. At present, the training categories of the demonstration platform is limited to the political, economic, military and such by using news corpora, but the built-in algorithm supports category customization training. The algorithm has high classification accuracy for conventional text, and the F value of the comprehensive open test is close to 86% (Gao et *al.*, 2017). There are two modes of text classification: *rule-based classification* and *machine learning based classification*.

Rule-based classification refers to the classification according to pre-defined classification rules, such as the “drugs” category, which can be further defined to be: “heroin; happy pills; poppy; cocaine; Pethidine; ecstasy; K powder”. The system will determine the text category according to the occurrence rules (words) in the text.

Machine learning classification is done through the use of an automatic machine learning implementation. Through a large amount of text training, the system has the ability to perform increasingly accurate classification. For example, if one prepares a large corpus of military and political categories, the classification effect is more and more accurate after continuous corpus training.

4.4.2. Text clustering

Text clustering can automatically analyze hot topics from large-scale data, and provide key feature descriptions of event topics. Text clustering is suitable for hot spot analysis of long texts, as well as short texts such as SMS and Weibo.

4.4.3 Reduplication detection

Document reduplication can quickly and accurately detect whether there are records of the same or similar content in the file collection or database, and find all duplicate records at the same time.

4.4.4 Information retrieval

The NLPIR-Parser use the JZSearch (or Precise Semantic Search) engine to support Full-text retirval. It supports text, numbers, dates, strings, and other data types. The query syntax supports AND/OR/NOT operations, NEAR proximity operations, and supports retrieval in Uighur, Tibetan, Mongolian, Arabic, Korean and other minority languages. It can be seamlessly integrated with existing text processing systems and database systems.

4.4.5 Intelligent filtering

Intelligent filtering can carry out semantic intelligent filtering and examination of text content, build complete thesauruses, intelligently identify a variety of variants (such as deformation, sound variation, complexity, and simplicity), and achieve semantic accurate disambiguation. The default system has 10 categories of nearly 40,000 keywords, and users can use “import keywords” to add personal keywords according to their needs. “Batch scan” can be used to filter bad information. Use “Open File” (or paste the text directly into the text box) for scanning.

5. Conclusion

As can be seen, compared with others analysis tools, such as WordSmith, AntConc, etc., the NLPIR- Parser is an enhanced semantic analysis platform. It can easily and effectively be used for the whole semantic analysis process for huge corpora without any further programming or any private data leaking; it can automatically identify new words and help generate all possible new term and domain specific lexicons if given unlabeled data; with the Chinese word segmentation tool and ICTCLAS integration, it supports automatic Chinese word segmentation; moreover, as it can present uniform tokenization and POS tagging for any mixture of English and Chinese texts, it is making Chinese and English semantic analysis easier and better.

Since 2000, the NLPIR-Parser has been released on the Natural Language Processing & Information Retrieval sharing platform (NLPIR.org). The application and source code can be accessed on Github (https://github.com/NLPIR-team/NLPIR). Until now, it has been licensed to over 400,000 entities all over the world, including Huawei, NewsCorp, NEC and Expert System. With its contributions to the semantic analysis field, the NLPIR-Parser has received numerous information Processing Awards, and is the tool-of-choice among academic and industry professionals in the constantly improving domain of semantic analysis.

Acknowledgements

References

Anthony L. (2005). AntConc: design and development of a freeware corpus analysis toolkit for the technical writing classroom. *International Professional Communication Conference*, IEEE.

Gao S., Zhang H.P. and Gao K. (2017). A Convolutional Neural Network Based Sentiment Classification and the Convolutional Kernel Representation. *Proceedings of 22nd International Conference on Natural Language & Information Systems* (NLDB), Springer LNCS, Liège, Belgium, 21-23 June.

Li R., Zhang H.P., Zhao Y.P. and Shang J.Y. (2014). Automatic Text Summarization Research Based on Topic Model and Information Entropy. *Computer Science*, vol. (11): 298-332.

Penn Treebank Project. (2003) Alphabetical list of part-of-speech tags used in the Penn Treebank Project, https://www.ling.upenn.edu/courses/Fall_2003/ling001/penn_treebank_pos.html

Salem A., Miao J. (2019). Le texte se transforme...Analyse textométrique des rapports d'ouverture présentés aux congrès du Parti Communiste Chinois (1982-2017). *HAL* Id:hal-02119927 https://hal.archives-ouvertes.fr/hal-02119927

Xi J.P. (2017). *Work Together to Build the Silk Road Economic Belt and The 21st Century Maritime Silk Road*. http://www.chinanews.com/gn/z/TheBeltandRoad/index.shtml

Xi J.P. (2019). *Working Together to Deliver a Brighter Future For Belt and Road Cooperation*. www.brfmc2019.cn

Yu S.W., Duan H. M., Zhu X.F., et al. (2002). Peking University Modern Chinese Corpus Construction (Cont.). *Journal of Chinese Information Processing*. 16(6).

Zhang H. P. and Shang J. Y. (2019). NLPIR-Parser: An intelligent semantic analysis toolkit for big data. *Corpus Linguistics*, 6(1): 87-104.

Zhang H.P. and Shang J.Y. (2017). Social media-oriented open domain new word detection. *Journal of Chinese Information Processing*. Vol (3): 115-121.

Zhang H.P. and Shang J.Y. (2019). *Big Data Intelligent Analysis*. Beijing: Tsinghua University Press.

Zhang H.P., Liu Q., Cheng X. Q., Zhang H.,Yu H.K. (2003). Chinese Lexical Analysis Using Hierarchical Hidden Markov Model, *Second SIGHAN workshop affiliated with 41st ACL;* Sapporo Japan, July, pp. 63-70.

Zhang H.P., Liu Q., Yu H.K, Cheng X.Q., Bai S. (2003). Chinese Name Entity Recognition Using Role Model. Special issue "Word Formation and Chinese Language processing" of the *International Journal of Computational Linguistics and Chinese Language Processing*, vol (8) : 29-602.

Tools

Anthony L. (2002), AntConc. Initially launched in 2002. The last version can be accessible in http://www.laurenceanthony.net/software.html

NLPIR. Parser. Github https://github.com/NLPIR-team/NLPIR/tree/master/NLPIR-Parser

Salem A. (1990/2018): *Lexico*, textometrical software developed at the ENS of St. Cloud and at the University of the Sorbonne nouvelle - Paris 3. Initially launched in 1990 and now version 5.0, http://lexi-co.com/.

Scott M . 2008. WordSmith Tools 5.0. Published by Lexical Analysis Software and Oxford University Press since 1996, now the version is 8.0. https://lexically.net/wordsmith/