Winter 8-2-2019

A Study on Adapting Natural Language Processing for Library Services Delivery

V Jeevitha
pinckjeevi@gmail.com

E.S. Kaviatha
kavithaesk@gmail.com

A Study on Adapting Natural Language Processing for Library Services Delivery

V. Jeevitha, Research Scholar, Department of Library and Information Science, Periyar University. pinckjeevi@gmail.com

Dr. E.S. Kavitha, Assistant Professor, Department of Library and Information Science, Periyar University. kavithaesk@gmail.com

Abstract

The technological revolution brought up with the advent of Internet demands that delivery of library services stay in tune with the technological advancements. Search technologies have evolved a long way from text based searching into voice recognized natural language queries. Technology Giants like Google, Apple and Microsoft have filled this marketplace with their products and services to tap the potentials. Integration of the Natural Language tools with existing library solutions provides an exciting potential for traditional libraries to expand their service base. For the users too, this provides a very meaningful experience to access the library resources in their natural language besides expanding into areas where traditional search mechanisms for resources may not lead into.

Key Words:

Natural Language Processing, Sentiment Analysis, Information Extraction, Name Entity Extraction, NLP in libraries, CoreNLP

Objectives

1. To study the need for adaption of technologies in the delivery of library services.
2. To study the key aspects and features of Natural Language Processing Tools and how their adaption will be good for libraries.
3. To suggest an interface for existing library solutions to integrate with NLP solutions for the delivery of library services
4. To provide a brief overview of the existing Open Source NLP Tools presently available and widely used in production.

Review of Literature

Gobinda G. Chowdhury (2000) Natural Language Processing (NLP) is an area of research and application that explores how computers can be used to understand and manipulate natural language text or speech to do useful things. NLP researchers aim to gather knowledge on how human beings understand and use language so that appropriate tools and techniques can be developed to make computer systems understand and manipulate natural languages to perform the desired tasks. The foundations of NLP lie in a number of disciplines, viz. computer and information sciences, linguistics, mathematics, electrical and electronic engineering, artificial intelligence and robotics, psychology, etc. Applications of NLP include a number of fields of studies, such as machine translation, natural language text processing and summarization, user interfaces, multilingual and cross language information retrieval (CLIR), speech recognition, artificial intelligence and expert systems, and so on.

Jeonghee Yi, Tetsuya Nasukawa, Razvan Bunescu & Wayne Niblack (2003) We present Sentiment Analyzer (SA) that extracts sentiment (or opinion) about a subject from online text documents. Instead of classifying the sentiment of an entire document about a subject, SA detects all references to the given subject, and determines sentiment in each of the references using natural language processing (NLP) techniques. Our sentiment analysis consists of 1) a topic specific feature term extraction, 2) sentiment extraction, and 3) (subject, sentiment) association by relationship analysis. SA utilizes two linguistic resources for the analysis: the sentiment lexicon and the sentiment pattern database. The performance of the algorithms was verified on online product review articles ("digital camera" and "music" reviews), and more general documents including general web pages and news articles.

MIT Press (2000) Increasingly businesses, government agencies and individuals are confronted with large amounts of text that are critical for working and living, but not well enough understood to get the enormous value out of them that they potentially hide. At the same time, the availability of large text corpora has changed the scientific approach to language in linguistics and cognitive science. Phenomena that were not detectable or seemed uninteresting in studying toy domains and individual sentences have moved into the center field of what is

considered important to explain. Whereas as recently as the early 1990s quantitative methods were seen as so inadequate for linguistics that an important textbook for mathematical linguistics did not cover them in any way, they are now increasingly seen as crucial for linguistic theory. In this book we have tried to achieve a balance between theory and practice, and between intuition and rigor.

Technological Advancements - Current Scenario

We live in a time where technologies are dynamically changing and fast eliminating their earlier siblings. Change has become the order of the day and anyone who doesn't innovate must necessarily hang his boot. This is prominently in service based industries/sectors. This rapid updating phenomena has created a user population whose demands too are constantly on the rise, and are easily unsatisfied with static versions of things which either fail to update preferring their constant mode, or their pace of updations are slow. While product developers or the service providers reel under the illusions that their products/services are free of any glitches and are offering best user experiences, the actual users will be just waiting for their time to switch to other similar products/services which not only offer their expected performance levels but also provide the latest state-of-the-art technological implementations. This indeed is a paradox, and library services too are no exempt from this.

Technology Adaption in Delivery of Library Services

Libraries have moved a long way from bin card based access control systems to the widely practiced OPAC Systems. Advancements like implementing Voice Interfaces for resources search, Multimedia Libraries, Special Purpose Libraries etc. have also been implemented. Libraries need to update and innovate their delivery services primarily to keep relevant in today's information avalanche. The normal user equipped with google and a smartphone is easily tempted to believe that all that he needs is at his fingertips with the input of some keystrokes rendering library a redundant place. Though apparently this may hold good, libraries contain treasure troves housing the cream of the finest pieces of human intelligence packaged with meticulous hard work, research and analysis. By ignoring to avail the library services, the individual easily forfeits all such opportunities that might have easily catapulted him upon proper utilizations. This must be averted both for the good cause of the user and the library.

The modern user is trendy and so too must be the library interfaces. Mere presence of vast collections is not a guarantee of their usages. Unless packaged with innovative interfaces, the modern user is more likely to skip the library, harming his potential for development besides causing an existential threat to the library.

NLP – The Innovatively Disruptive Technology

One such innovative technology could be to use Natural Language interfaces in the deployment Library Services. In simple terms sans jargons, Natural Language interfaces are those kinds of interfaces whereby the user simply interacts with the computer as though he is involved in a verbal conversation with another human being. Technology majors like Google, Apple and Microsoft have already realized the potential of Natural Language interfaces and have implemented the same in their existing products and services, as OK Google, Siri and Cortana respectively. By similar adaption of Natural Language Interface for the delivery of library services, the libraries can induce and ensure that more and more number of users avail the library services.

Key Aspects and Features of NLP Systems

Let us now throw some light on the major aspects and features of what constitutes an NLP system.

Natural Language Understanding

In plain language, Natural Language Understanding or NLU means decoding the input in human language into meaningful machine representations. This involves storage of pre-constructed representations for human language, segregating the human input into individual representations and mapping the segregated input against them.

Natural Language Generation

Since NLP responds to the user in human language only, the NLP system respond for the abstract output of the query raised by the user in human understandable language and put it in phrases and sentences in the human language.

Information Extraction

Information Extraction refers to the process by which structured information is extracted automatically from unstructured and/or semi-structured sources, such as text documents or web pages for example. Rapid advancements made in Artificial Intelligence has made it possible to automatically decipher information stored in repositories like documents, databases, articles, books etc. and make sense out of that information by identifying and grouping them under special ids, also called as tags. For example an NLP system might scan through the list of given PDF files that narrate the famines or the great achievements made in each field of science and group them under different tags so that whenever any reference is made later to any of the associated tag, the NLP system will automatically be able to correlate the same with its stored entities and present the detailed list to the user.

Named Entity Extraction

While analyzing a document, the NLP systems locate and classify elements contained in the text of the document into predefined categories such as the names of people, organizations, places, monetary values, percentages, achievements related to specific title, for example, long jump track records, runs scored by each player in a cricket match etc. This feature s called Named Entity Extraction. Through this feature, the NLP systems are able to respond correctly to user queries by classifying the user queries into elements and mapping them to the stored entity list.

Sentiment Analysis

Sentiment Analysis is the most important feature of any good NLP system. Also known as opinion mining , it is the process of determining the emotional tone behind a series of words, used to gain an understanding of the the attitudes, opinions and emotions expressed within an article or an online mention. Sentiment analysis extracts subjective information in source material.

Word Sense Disambiguation

Word Sense Disambiguation refers to the process through which the context sensitive meaning of the word/phrase is performed by the system. For example, in a system related to

Blood Pressure, the word "High" may refer to something critical, whereas in the system related to student performances the very same word "High" may relate to something very good.

Explicit Semantic Analysis (ESA)

Explicit Semantic Analysis (ESA) is the process of understanding the meaning of a piece text, as a combination of the concepts found in that text. To understand ESA clearly, let us discuss about document classification. In document classification, documents are tagged to make them easier to manage and sort. Tagging a document with keywords makes it easier to find. However, keyword tagging alone has its limitations; searches carried out using vocabulary with similar meaning, but different actual words may not uncover relevant documents. Besides, document tagging is a serious process and requires continuous updations and improvisations. Any miss may result in the likelihood of the document being omitted even though it may actually contain what the user wants. These drawbacks are permanently eliminated using ESA. By classifying text semantically i.e. representing the document as concepts and lowering the dependence on specific keywords can greatly improve a machine's understanding of text.

NLP in Action in a Library – An Imaginary Scenario

An Example Scenario of an imaginary conversation between the library user and the library solution implementing natural language interface in book lending services of the library is given below;

Library User: Hey, can you help me out?

Library Solution: Yes, Sure, Please tell me what do you need?

L U: Is the book Selfish Gene available?

L S: Please wait. I will come back to you.

L S: Do you mean The Selfish Gene by Richard Dawkins?

L U: Yes

L S: We have it. Presently it is out on lending and will be returned by 30th of this month, should the current user choose not to renew it.

L U: Thanks for the response.

L S: Thank You.

Technical Dissemination of the conversation

Let us now make a map of the conversation with the terms discussed above. Each query by the Library User requires the system to employ NLU. The output from the NLU is then fed to the Information Extraction functionality whereby the tokenized query of the user is analysed and information contained in it is extracted. The Named Entities present in the query, namely the book or author or title or publication are also extracted from the input. Word Sense Disambiguation is also applied so as to understand the contextual meaning of the terms present in the user query. Explicit Semantic Analysis (ESA) is applied by understanding the context of the query and the response map is prepared by the system working in tandem with the library solution. Not merely the specific Boolean answers of available/not available for the query of the user, but rather a comprehensive detail of its availability status and its return date are estimated from the library database and then a structured response in human language is generated by the Natural Language Generation portion of the NLP and presented to the user. Also when the user input doesn't contain any query portion, the NLP understands it as the termination of the query and responds courteously.

The above conversation though sounds bit far stretching imagination, is a distinct possibility and it is only a matter of time before libraries make it a reality.

Integrating Library Solutions with NLP interfaces in Libraries – Schematic Diagram
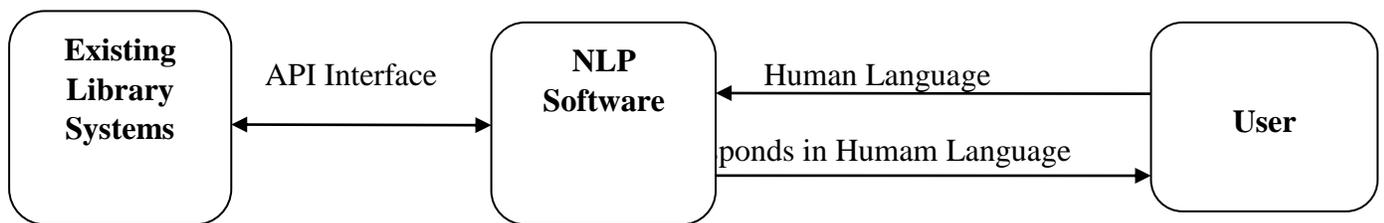
As seen from the above diagram, NLP typically acts as a middleware tool. The existing library solutions are interfaced with the NLP tool of choice using Application Programming Interfaces on one hand. The NLP tool in turn interacts with the user, disseminates his natural query, sends them to the library solution and gets the response and remits the same to the user in natural language.

Tools and Strategies to integrate Library Solutions with NLP

Though a lot of tools are available in the open market for NLP, we will discuss the most prominent tools, being CoreNLP and NLTK only. It has to be noted that both CoreNLP and NLTK are Open Source, meaning that they can be used at free cost by many users for non proprietary purposes.

CoreNLP

CoreNLP from Stanford University provides a comprehensive set of human language technology tools. It gives the base forms of words, their parts of speech, whether they are names of companies, people, etc., normalize dates, times, and numeric quantities, mark up the structure of sentences in terms of phrases and syntactic dependencies, indicate which noun phrases refer to the same entities, indicate sentiment, extract particular or open-class relations between entity mentions, get the quotes people said, etc. CoreNLP is fast, robust and widely used in production environments. It has available Application Programming Interfaces in major modern programming languages, which means that its integration into existing solutions is easier. To integrate Library Solutions with CoreNLP, the librarian along with the Software Professionals can create working models which can then be taken by the development team for actual implementation.

Natural Language Tool Kit

NLTK or Natural Language Tool Kit, is a leading platform for building programs in Python Programming Language to work with human language data. It provides easy to use interfaces and includes almost all the aspects of NLP systems. It has an API which can be used to interface the existing solutions with NLTK. Like CoreNLP, NLTK too promises exciting new interfaces for innovative library services delivery.

TextBlob, GenSim, SpaCy are some of the other tools widely used in NLP. Besides these, a whole bunch of proprietary and custom built NLP tools are also available.

Conclusion

The NLP integration with library solutions is poised for big growth in the days to come. NLP with its capabilities to expand the horizons of search breaking the barriers imposed by the

current technologies will come in handy for library users. Once adapted, library users will find quality contents returned by NLP perfectly fitting the intent of their search, which will result in better utilization efficiencies of library resources. Because of its ability to read mine through the text and ability to ascertain the sentiment behind the expressions, the NLP solutions will greatly reduce the laborious and tedious processes undergone by the users presently. This will be a major incentivizing factor for library utilizations besides natural language interfaces. Thus we can safely conclude that inclusion of NLP services will augur well for libraries and present a new dimension in the delivery of library services.

Reference

1. https://data-flair.training/blogs/natural-language-processing/
2. https://blog.algorithmia.com/introduction-natural-language-processing-nlp/
3. https://searchenterpriseai.techtarget.com/definition/natural-language-understanding-NLU
4. https://towardsdatascience.com/5-heroic-tools-for-natural-language-processing-7f3c1f8fc9f0
5. http://blog.aylien.com/10-common-nlp-terms-explained-for-the-text/
6. https://www.springer.com/in/book/9783642047343
7. https://machinelearningmastery.com/applications-of-deep-learning-for-natural-language-processing/
8. https://www.techemergence.com/nlp-current-applications-and-future-possibilities/
9. https://strathprints.strath.ac.uk/2611/1/strathprints002611.pdf
10. https://dl.acm.org/citation.cfm?id=952133