ISSN 1799-2591
Theory and Practice in Language Studies, Vol. 10, No. 1, pp. 49-54, January 2020
DOI: http://dx.doi.org/10.17507/tpls.1001.07

# Overview of Natural Language Processing Technologies and Rationales in Application*

Fei Song
Beijing International Studies University, China

Jun Sun
Beijing International Studies University, China

Tao Wang
Beijing International Studies University, China

***Abstract*—In the past decade, rapid advancement of new technologies including data technology, virtual reality (VR) and artificial intelligence (AI), which are all related to language disciplines, brings a new era of data-based language studies, relying on AI to enhance the language ability and VI to create fresh new experience. Practice of language processing in language disciplines by those technologies in turn promotes the emergence of some other revolutionary technologies, for example, the increasingly common data thinking and computational thinking in language research. In this context, it is of great significance to seize the opportunity of big data era, and make full use of AI and other new technologies to substantially promote language-related studies. Thus, an overview of several important language processing technologies and the corresponding rationales, as well as the latest progress is expounded in this paper.**

***Index Terms*—natural language processing technologies, data thinking, computational thinking, overview**

## I. Language Processing and Technology

Language processing, generally referred to as Natural Language Processing (NLP), is a way to study the theory and methods of effective communication between humans and machinery. For instance, NLP can be regarded as a process to teach computers to learn human natural language. Though belong to different fields, language processing and language teaching actually share deep-rooted similarities, where NLP simulates the cognitive characteristics of human beings in language learning and use in a statistical language model, and the practice of NLP helps to uncover the laws of language teaching (Song Fei 2018), and thus, NLP can be subdivided into natural language understanding (NLU) and natural language generation on the basis of the functions of human brain to process language.

In this paper, instead of elaborating in strict accordance with NLP disciplinary framework, specific technologies closely related to people's life and breakthrough applications in recent years are introduced, to facilitate the understanding for those without the background of science and engineering.

## II. Natural Language Understanding Technology (NLUT)

In a narrow sense, NLU does not include speech recognition and characters recognition. However, in a broad sense, any technology involved in making computers "understand" human languages can be included into the field of NLU, of which the latter is adopted in this paper. Over the years, the NLUT, which is closely connected with people's life, mainly involves information retrieval, text clustering, speech recognition, characters recognition, affective computing and other fields. It is not intended to cover too much of the apparent application of these technologies in this paper (after all, living in the information era, people cannot have no idea of them), but aims to present the rationales behind these seemingly "intricate" technologies with plain expressions and examples.

### A. Information Retrieval (IR)

IR is not a new word; and its related technology is indispensable in people's life today. Nevertheless, in the modern business model, the search engine, closely related to IR technology, just came up at the end of the 20th century. Currently, Google can be treated as the unicorn among those companies started from IR technology in the world. Since co-founded by Larry Page and Sergey Brin in 1998, Google's industrial chain has extended from search engine to hardware (Chrome Book Notebook, Nexus Mobile Phone), virtual reality (Google Glass), biological technology (Calico), smart home (Nest) and other fields.

Among the numerous algorithms involved in Google search engine, TF-IDF, which solves the problem of measuring

* It is supported by The Social Science Foundation of Beijing (Project No.: 16YYC028)

the relevance of web pages and search terms, plays a decisive role. From the perspective of web pages, the higher the frequency of search words in a web page, the more relevant the web page is to search, which is so-called TF (Term Frequency). In terms of the difference of the importance for each search word, a retrieved word can have a stronger ability to locate the web page if it appears in only a few web pages, because of less non-target web pages; and vice versa, another retrieved word may have much weaker ability to locate the web page if it appears in numerous web pages, which is so-called IDF (Inverse Document Frequency). The calculation formula of IDF is as follows:

IDF=log（D/Dw）

Of which, D is the total web pages, w the retrieved word, and Dw the number of web pages appearing the retrieved word. The specific mechanism is to assign values to the ability of different retrieved words to locate web pages. For example, a user inputs "太阳能应用" for retrieval, assuming that the total number of web pages is 2 billion, and that the retrieved word "太阳能" appears in one million Web pages, then, its IDF is log (2 billion / 1 million), namely 11.0. Meanwhile, "应用" has appeared in one billion web pages, and its IDF is log (2 billion / 1 billion), which is 0.7. For this reason, "太阳能" contributes as much toward locking down web pages as 16 "应用" do, which is more in line with people's intuitive perception.

In addition to TF-IDF, PageRank is another Google's key core technology, which solves the problem of page ranking in information retrieval results. Through machine retrieval, it is not difficult to hit the data containing the retrieval words, but how to prioritize thousands of retrieval results is of vital importance. After the emergence of PageRank technology, the ranking relevance of search results undergoes a qualitative leap, thus establishing Google's dominant position in the field of search engines. As is shown in its name, the technology is developed by its founder (Page et al.). In spite of its great significance, the basic principles of NLU involved are uncomplicated at all.

If "马云" is searched, after checking the public security system, 10 thousand "Jack 马云" will appear, for example. However, which one is the person looking for? If everyone says that Jack Ma of Alibaba is authentic, then it surely is. Therefore, the principle can be summarized as the following two aspects: first, the more links a web page is linked to by others (more inbound links), the higher the degree of trust is, so it is with its ranking; second, the links provided by the top ranked pages are more important than those by the low ranked ones, and the same goes for the weight.

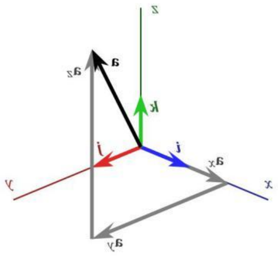
In China, two search engine companies, Google and Baidu, coexisted years ago, until Google withdrew from China due to legal issues. The withdrawal was interpreted by many foreign media as "force-out", which is rather misconceived. Nonetheless, the search engine, from another perspective, based on information retrieval technology is related to the big data problem of Internet users nationwide, which is of great significance to the national network information security. To this point, a search engine company cannot survive in these places by violating the laws and regulations there. On August 2019, the high-tech company "ByteDance" announced that it would conduct a full web search, which is expected to challenge the dominance of Baidu in China's current search engine industry.

### *B. Text Clustering*

According to the clustering hypothesis, the similarity of homogeneous documents is larger than that of inhomogeneous ones. Thus, merging the homogeneous documents is called text clustering. It seems that the cosine theorem and the merging of homogeneous documents are two things related to one another as an apple to an oyster, but these two have exactly produced magical chemical reactions.

The essential problem to be solved in classifying articles lies in how to measure the similarity among articles. Apart from those subjective feelings, there is also a quantitative comparison method for the similarity between the two articles, namely, transforming an article into a vector quantity with direction and length, which then can show difference between them, after calculating the included angle between the two articles with the cosine theorem. The remaining task is how to turn an article into a vector quantity.

When an article is set as a feature vector, it can be composed of multi-dimensional component vectors representing all the words possible showing in all articles. To ensure that component vectors are the same, taking the same dictionary as an example, if the number of words received is 80, 000, then each article can be expressed as a total vector formed by adding the 80, 000- dimensional component vectors. In the article, some words are more important for the classification of articles, while others are less important. Intuitively, the function words like "的", "了", "得" seem unimportant, but by the words "股票, 血小板, 投篮", it seems easier to distinguish the theme, precisely corresponding to the IDF mentioned above. On top of that, the high-frequency words in an article are usually more conducive to classification than the low-frequency ones. Therefore, it is necessary to calculate the specific length of the 80, 000 component vectors in each article, which exactly corresponds to the TF mentioned above. It will be thus seen that, each article can be mapped to a total vector (Feature Vector), and the size of each dimension in the vector represents the contribution of each word to the classification of this article. When articles are transformed into feature vectors, then the included angle (similarity) between them can be calculated.



Different articles have different length, which means their length of the feature vectors in each dimension is naturally different. This sort of length comparison offers no help to better compare the similarity of articles. However, the included angle between vectors is all that matters. The included angle can be calculated according to the cosine theorem.

Suppose that the TF-IDF values corresponding to the words in the two articles X and Y are x UU 1, X UU 2, · · · · · · · · · · · , X and Y1, Y2, · · · · · , y80000, then the cosine of the included angle between them is:

$$\cos \square = \frac{x_1 y_1 + x_2 y_2 + \cdots + x_{80000} y_{80000}}{\sqrt{x_1^2 + x_2^2 + \cdots + x_{80000}^2} \cdot \sqrt{y_1^2 + y_2^2 + \cdots + y_{80000}^2}}$$

Thus, the similarity between the two articles is transformed into a specific value. After the threshold value is set and iterated upward continuously, the category will be on the decrease, while the number of articles in this category is growing and the similarity of articles is reducing. If the similarity lowers than a certain degree, larger categories will no longer be merged, and then the text categorization is completed.

Text clustering technology is often used to categorize topics of news. After that, automatic abstracts can be further generated, thus realizing the automatic collecting and editing of news.

### C. *Speech Recognition*

Speech recognition technology currently enjoys wide application scenarios, like the fields of shorthand, automatic question and answering system, map navigation and others. One of the most relevant aspects in ordinary people's life is probably the voice-to-text function in WeChat. Speech recognition is now an indispensable part of artificial intelligence, and the basic principles behind the seemingly profound appearance are not complicated at all.

A sentence contains many words, and each word will have several homonyms, which means many possible combinations for this sentence, so speech recognition needs to figure out the most likely combination of words through calculation from a great number of combinations.

P(S)=P（w1,w2,…,wn）=P（w1）·P（w2| w1）·P（w3| w1,w2）…P（wn| w1,w2,…,wn-1）

Suppose S is a sentence with a specific meaning, which consists of a group of words w1, w2, …, wn, arranged in a particular order. Possibility of sentence S in natural language, is also the probability P (S) that needs to be worked out. Expand S, it can be found out that:

P(S)=P（w1,w2,…,wn）=P（w1）·P（w2| w1）·P（w3| w1,w2）…P（wn| w1,w2,…,wn-1）

Of which, P (w1) is the probability of the first word, and P (w2| w1) the second word under the premise of the first word, also known as the conditional probability of the second word. The rest may be inferred that, P (wn| w1, w2, ..., wn-1) is the probability of the last word after all the previous words appear. Since the value space of each variable w is the size of a dictionary, the calculation of conditional probability will be more complicated. To simplify the operation, the Russian mathematician Andrey Markov put forward the Bigram Model, namely, suppose that the probability of each word is only related to the word that precedes it. The facts proved that the Bigram Model is far enough to solve many practical problems. In the simplified Bigram Model, the probability P(S) of sentence S is calculated as follows:

P(S)=P（w1,w2,…,wn）=P（w1）·P（w2| w1）·P（w3| w2）…P（wn| wn-1）

The next is to calculate the conditional probability P（wi|wi-1）to figure out P(S). According to the definition of the conditional probability:

P（wi|wi-1）= $\frac{\mathrm{P}(\mathrm{w_i},\mathrm{w_{i-1}})}{\mathrm{P}(\mathrm{w_{i-1}})}$

It is not difficult to evaluate the marginal probability P $(\mathrm{w_{i-1}})$ and the joint probability P $(\mathrm{w_i}, \mathrm{w_{i-1}})$ , and only by collecting the on-demand corpus and establishing a corpus or balanced corpus in the corresponding field that meets the requirements of the language model in the computer, can the frequency of words and the frequency of any two word collocations be calculated by computer. If the corpus is large enough and properly matched, the frequency can be regarded as probability approximately. The marginal probability $\mathrm{P}(\mathrm{w_{i-1}})$ can be retrieved from the word frequency database, while the joint probability $\mathrm{P}(\mathrm{w_i}, \mathrm{w_{i-1}})$ from the collocation frequency database. From the things mentioned, the probability of any sentence in natural language can be calculated.

Another example for those without mathematical basis to understand, is that an author voices "wǒ shì yī gè zhōng guó rén" to Siri. When the server receives this series of pronunciations, it will first retrieve the first syllable to see which word has the highest frequency among all Chinese words pronounced "wǒ". As the retrieval results show, the four words pronounced "wǒ" and their word frequency data are as follows:

TABLE
CHINESE WORDS PRONOUNCING "WǑ" AND THEIR WORD FREQUENCY

| Pronunciation | wǒ | | | |
|---|---|---|---|---|
| Chinese characters | 我 | 婑 | 媒 | 捰 |
| Word Frequency | 115623 | 5 | 2 | 3 |

As is shown in the Table above, the frequency "我" is the highest among them, so the server assumes that the first word is "我". Afterwards, the second syllable "shì" is retrieved and then all the words pronounced "shì" are retrieved. Next up, the co-occurrence frequency of "wo (我)" and these words is found to be the highest, and the server assumes that the second word is “shi”. Similarly, the server combines all the possible words of all the syllables in this sentence, then figures out the probability of each possible sequence to find the one with the highest probability, and finally identifies the sentence that the author has said (Song Fei 2018).

Currently, iFLYTEK, a Chinese company, is in the leading position in voice recognition technology worldwide, launching a series of important products and services based on speech recognition, such as iFLYREC, iFLYTEK Easytrans, etc. In addition, “SoGou” Company also launches “SoGou Smart Recorder”, which can realize timely conversion of recording based on speech recognition.

Nowadays, more and more intelligent devices based on speech recognition technology have entered people’s home. For instance, the popular intelligent speaker in the recent two years has applied the technology of speech recognition and wake-on-voice, bringing a lot of joy to people’s life.

### *D. Words Recognition*

Technically, words recognition cannot be classified into the category of natural language understanding (NLU) technology, because its core technology applied should belong to image recognition. However, since it involves text and is also a writing symbol that helps the machine “understand” the human language in a broad sense, it will be briefly introduced here. Words recognition technology is often used in some PDF document reading editors, such as Adobe Acrobat, CAJ Viewer, and so on, which is often seen in the software as a button, that is, “OCR” (Optimal Characters Recognition). Generally speaking, the PDF file obtained by scanning is essentially the same as the ordinary picture and the word is only normal image with optical characteristics. It cannot be directly extracted as text by text editing software (such as MS Word). At this point, the words recognition technology is required to identify and extract word in a file. Thus, the recognized words can be directly extracted and edited by the word editing software.

IFLYTEK has achieved certain results in Handwriting Words Recognition. This technology is being applied to fields like data archiving and assisted instruction.

In addition to the simple and traditional Chinese characters commonly used today, words recognition technology is being applied to the recognition of ancient writing. In May 2019, the Chinese Character Research and Application Center of East China Normal University (ECNU) released the “AI+ Ideogram Big Data Achievement - Smartscope for Characters Used in Dynasties of Shang, Zhou and Jin”, which is an attempt to identify ancient characters by using words recognition technology.

### *E. Affective Computing*

Affective computing, also known as “sentiment analysis”, is a field involving a variety of high-tech. The main goal is to simulate human emotions with the assistance of AI. According to the analysis, affective computing can be speech-based, text-based, expression-based, physiological-based and others, of which the latter two are not discussed here because they do not involve language.

Speech-based affective computing mainly realizes the understanding and simulation of human affection by means of speech features, such as short-term energy and short-term average amplitude, pitch period, short-term zero-crossing rate, speech rate and so on. Text-based affective computing, mainly through lexical, grammatical and other language elements to achieve deep semantic analysis involving emotions, is one of the important contents of network public opinion analysis. At present, it is those social medias (such as “Sina Weibo”) that adopt text-based affective computing in China. By crawling large-scale automatic user data, the corpus is built and then processed through text processing such as automatic segmentation. Finally, a specific algorithm is used to analyze the user’s affection (emotion) .

## III. Natural Language Generation Technology

Similar to natural language understanding, natural language generation, in a narrow sense, means to enable computers to possess the same function of expression and writing as human beings, mainly referring to text here. And a broad sense, the technology involved in having the computer “generate” the human language can be considered as the field of natural language generation. Speech can be viewed as a medium of language, so generating speech also means generating human language. This section will mainly focus on speech synthesis and machine writing.

### *A. Speech Synthesis*

Speech synthesis, can be generally regarded as the employment of computers and electronic devices to simulate the generation of human speech, which has undergone such phase as parameter synthesis and waveform stitching. In some cases, speech synthesis technology is limited to “text-to-speech” (TTS) technology, and often applied in AI-based customer service, text reading software, mobile phone ring tones, and the like.

Some may have a deep impression on the voice prompts of the bus reporting stations in previous years. In the voice prompts, the combination of words and words is usually unnatural, and the speed of speech is not balanced, obviously sounding unlike a real person. However, to some extent, this voice prompt is also a technique that speech synthesis used. In addition, many people will imitate the robot’s speech word by word, and will also use the intermittent movement of the body’s joints to mimic the movement of the robot back at childhood. In fact, nowadays, the voices that robots can make, or the actions they can made, are not the ones that people imagined twenty or thirty years ago, but they are very coherent and smooth, and not much different from real people. In this field, domestic companies such as “IFLYTEK” are at the forefront.

### *B. Machine Writing*

Machine writing refers to the technology that computers write articles on specific themes and with related materials. At present, there are two main technical routes for machine writing: one is to use a template to fill specific data and information into the corresponding position of the template to generate an article; the other is to obtain a large amount of data, integrate the information in the data, and reorganize the output language. Nowadays, machine writing is mainly used for the creation of news reports. Through the rapid integration of data, a large number of documents can be generated in a short period of time to ensure the timeliness of news. In additions, there are robot poets invented, such as Microsoft “Xiaoice” who can even “compose poems from pictures”, and in 2017 published the first collection of robotic poems Sunshine Misses Windows. With the technical advancement, there are even robot journalists (such as “Cat AI”, Giiso, etc.) that help self-editing media articles for re-processing. In addition, machine writing also facilitates the development of automatic abstracting technology.

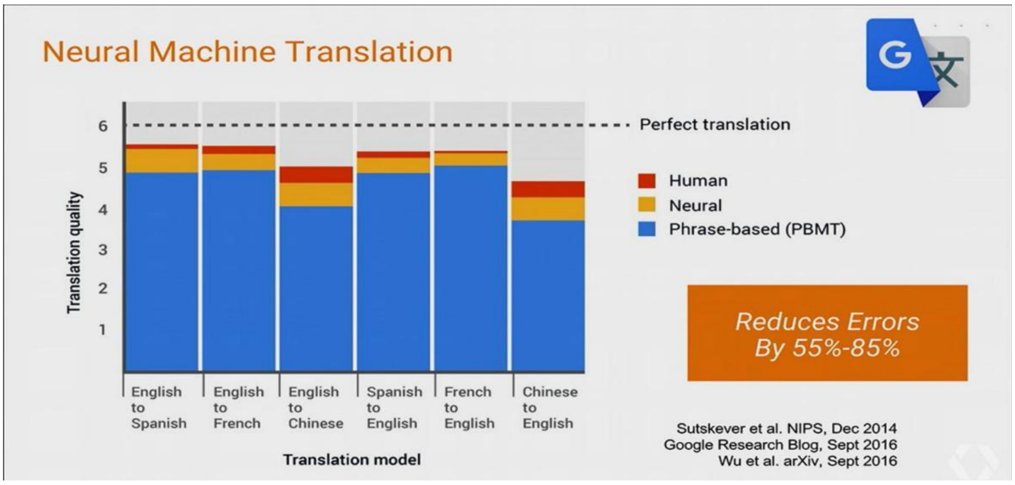
Compared with human writing, the current machine writing is advantageous in quantity and efficiency, but lacking in originality and connotation.

## IV. The Intersection of Natural Language Understanding and Generation

### *A. Machine Translation*

Machine translation, regarded almost as one of the earliest prospects of natural language processing or AI, has been considered and experimented as early as the computer appeared in the 1940s and 1950s. It contains both natural language understanding techniques and natural language generation, because the computer must first understand the source language, and then generate the same semantics as the source text in the target language.

However, it is not until the 1980s that machine translation can make substantial progress, mainly because of the limited performance of computers at that time, which makes it difficult to perform efficient calculations for a large amount of data, and also restricts the researchers’ researches on rules-based (grammar) machine translation. In the late 1980s, the emergence of microprocessors enables computer capabilities with leapfrog development, and hence the potential and economic benefits of machine translation discipline can be vitalized. In the meantime, some basic research in computational linguistics, such as many important algorithms and research on grammar and semantics, have achieved some important results. Later, with the continuously increasing performance of computers, the researchers turn their eyes to statistical-based researches. Nevertheless, the initial results are still unsatisfactory until the beginning of the 21st century, when then there was a popular “punchline”; that is, the sentence “How old are you?” was translated as the meaning of “Why are you again?” by machine translation.



Neural Machine Translation

Reduces Errors By 55%-85%

Sutskever et al. NIPS, Dec 2014
Google Research Blog, Sept 2016
Wu et al. arXiv, Sept 2016

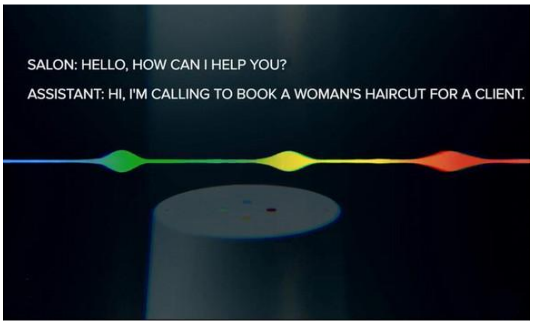
However, the development of machine translation in the past decade or so can be said to be earth-shaking. In a statistical method, machine translation has gone from “word alignment” to “phrase alignment” and then to “sentence alignment” based on neural network, with unprecedented accuracy. At the end of 2016, Google Translate developed and used the Google Neuro Machine Translation (GNMT, Google NMT), a formal application of neural network machine translation. Compared with the previous statistical models, neural network machine translation is smooth translation, accurate, understandable and fast-speeded.

At present, machine translation has been widely applied in the translation industry, reduce the heavy workload of translators. In addition, handheld “translators” have emerged to meet the basic communication needs for tourists, playing a role in many fields of society.

### *B. Automatic Question and Answer (Q&A)*

Speaking of “human-machine dialogue”, it often occurs to people in a scene of a person talking with a robot. Narrowly speaking, the concept of “human-machine dialogue” should be an automatic Q&A technique.

The automatic Q&A system and voice recognition and speech synthesis have all together been applied into the intelligent customer service. Of course, in addition to that, it includes companion robots, smart speakers and other products.



SALON: HELLO, HOW CAN I HELP YOU?

ASSISTANT: HI, I'M CALLING TO BOOK A WOMAN'S HAIRCUT FOR A CLIENT.

The famous “Turing Test”, in fact, is based on an automatic Q&A system. In the I/O Developers Conference held in 2018, Google demonstrated the impressive Duplex AI voice technology, which can simulate human tone, speech rate, and help users reserve hair salons and restaurants in a smooth human-computer interaction. On the last day of the conference, John Hennessy, chairman of Google’s parent company Alphabet and former Stanford president, announced that Duplex had passed the Turing test.

## V. Conclusion

From this point of view, technology and language processing have actually been interlinked and integrated with each other nowadays; namely, computer can facilitate language processing which in turn applies many cutting-edge computer technologies of human beings especially in the field of AI for language studies, particularly language teaching.

## References

[1] Shi Jie. (2019). The Status Quo and Future of Artificial Intelligence to Aid Foreign Language Teaching. Beijing International Studies University, 2019.

[2] Song Fei. (2018). The Common Essence of Second Language Teaching and Natural Language Processing From a Statistical Perspective. *International Chinese Language Education (Chinese and English), 3*, 39-45.

[3] Song Fei, Wang Yadan, Lan Qingqing, Shi Jie. (2018). New Technology Application Development Report in Chinese International Education. Chinese International Education Development Report (2015-2016). Beijing: Social Sciences Academic Press (CHINA).

[4] Ye Weibing, Liu Shijuan, Song Fei. (2017). History and Current State of Virtual Reality Technology and Its Application in Language Education. *Journal of Technology and Chinese Language Teaching, 2*, 70-100.

[5] Wu Jun. (2014). The Beauty of Mathematics. Beijing: Posts & Telecom Press.

**Fei Song** was born in Linyi, China in 1986. He received his PH.D. degree in linguistics from Minzu University of China in 2015.

He is currently the associate professor of Beijing International Studies University. He focuses on Chinese language processing and international Chinese teaching.

**Jun Sun** received Master degree from Beijing Language and Culture University. He is an associate professor of Beijing International Studies University. And he is specializing in International Chinese Teaching and Intercultural Communication.

**Tao Wang** got Master degree from Nanjing University. He is now a lecturer of Beijing International Studies University. His research field is international Chinese Teaching and technology and Chinese teaching and he has published the Chinese-language micro-lens.