ISSN No. 0976-5697

Volume 8, No. 3, March – April 2017

International Journal of Advanced Research in Computer Science

REVIEW ATICLE

Available Online at www.ijarcs.info

Developments in The Field of Natural Language Processing

Bhargavi Goel
Delhi Public School, Vasant Kunj
Delhi, India

Abstract: Natural language processing involves computer science, artificial intelligence and computational linguistics concerned with interactions between computers and human (natural) languages. The paper attempts to critically analyse state of the art technology algorithms in the field of Information Extraction and Information Retrieval. Information Extraction is concerned in general with the extraction of semantic information from text. Retrieval, filtering, indexing and other such tools have been built which have been used to accomplish tasks such as named entity recognition, co-reference resolution, relationship extraction, etc.

By collating important work systematically, the paper also aims to simplify the process of referencing and literature review for future researchers and developers in the field of Natural Language Processing. Major challenges in NLP including natural language understanding, enabling computers to derive meaning from human or natural language input; natural language generation among others have also been discussed.

Keywords: Natural Language Processing, Information Extraction, Information Retrieval, Machine Translation, Natural Language Generation.

I. INTRODUCTION

Natural Language Processing (NLP) is a field of research which is focused on the exploration of the usage of computers in understanding and manipulating natural language text or speech for useful purposes. It can also be described as a computerized approach for the analysis of text on the basis of a set of theories and technologies. It has been defined by Elizabeth D. Liddy (2001) as: “A theoretically motivated range of computational techniques for analyzing and representing naturally occurring texts at one or more levels of linguistic analysis for the purpose of achieving human-like language processing for a range of tasks or applications.”

NLP researchers have been placing emphasis on the collection of knowledge about the understanding of human beings and the use of language for the development of appropriate tools and techniques to allow the understanding and manipulation of natural languages by computer systems for the performance of desired tasks. NLP is believed to have originated from a wide range of disciplines including computer and information sciences, linguistics, mathematics, electrical and electronic engineering, artificial intelligence and robotics and psychology.

II. BRIEF HISTORY OF NLP

The field of NLP emerged in the late 1940s and research has continuously been carried out on it since then with commendable advancements in the last few decades. The earliest phase of interest in NLP can be marked with the publication of a paper entitled ‘Computing Machinery and Intelligence’ in 1950 by Alan Turing. This period also saw the emergence of the Turing test. His work asserted the importance of intelligent computers that could carry on conversations with humans without them realizing that they were talking to machines. His work further led to discoveries and developments over the subsequent decades which have brought about major shifts in the discipline of Natural Language Processing.

A. *The era from late 1960s-1970s*

The issue of representation of meaning and the development of computationally tractable solutions was not producible by the then-existing theories of grammar marked by the theoretical work in the late 1960’s and early 1970’s. The transformational model of linguistic competence was introduced in 1965 by Chomsky (Chomsky, 1965), a flagship development in the field of NLP as it marked the beginning of the era of NLP research. However, it had the shortcomings of the transformational generative grammars being too syntactically oriented to allow for semantic concerns in addition to being difficult for computational implementation. In response to the theories of Chomsky (1956, 1965) and the work of other transformational generativists, many theories were proposed to describe the syntactic anomalies, and provide semantic representations including case grammar of Fillmore (Fillmore, 1968), semantic networks of Quillian (1968), and conceptual dependency theory of Schank (1975). In addition to these, the power of phase structure grammar was extended by Woods (1970) through his augmented transition networks which also incorporated mechanisms from programming languages such as LISP. Wilks’ preference semantics (1975), and Kay’s functional grammar (1979) were the other representation formalisms of that era.

In addition to theoretical development, the period also saw the development of many prototype systems to demonstrate the effectiveness of particular principles including Weizenbaum’s ELIZA (1966) which was designed to echo user input and cause the replication of the conversation between a psychologist and a patient. Other than this SHRDLU by Winograd (1971) and LUNAR by Woods (1970) were demonstrations of the use of NLP in the development of operational systems with real levels of linguistic processing, in truly end-to-end (although toy) systems. Although these were not able to accomplish major goals, they still managed to inspire the development of more complex systems of processing in real world systems. PARRY (1972), for example, was inspired from Winograd’s SHRDLU.

A shift in attention to semantic issues, discourse phenomena, and communicative goals and plans was observed in the late 1970s. This could be symbolized by the work of Grosz (1979) who analyzed task-oriented dialogues and proposed a theory to partition the discourse into units, and Mann and Thompson (1988) who developed Rhetorical Structure Theory, attributing hierarchical structure to discourse.

Other researchers who have contributed significantly to this field include Hobbs and Rosenschein, Polanyi and Schank, and Reichman. Considerable advancements in natural language generation were made in this period as well which can be exemplified by McKeown's discourse planner TEXT which came with the ability to generate coherent responses online and McDonald's response generator MUMMBLE.

B. *The 1980s*

In the early 1980s, non-symbolic approaches again became the popular area for research due to the availability of critical computational resources resulting from Moore's law and the growing awareness of the limitations of isolated solutions to NLP problems. This resulted in an increased impact of these statistical approaches by the end of 1980s, bringing them at par with the symbolic approaches which were based on complex sets of hand-written rules. The revolution in NLP began with the introduction of machine learning algorithms for language processing like decision trees, which were one of the earliest-used machine learning algorithms and produced systems of hard if-then rules. The research then shifted to more reliable models which could be integrated into larger systems consisting of multiple subtasks and making soft decisions on the basis of input data. The cache language models upon which many speech recognition systems now rely are examples of such statistical models.

Development of algorithms allowing for a program to infer patterns about sample data to make predictions about new data was required for the statistical and machine learning models. This requires an iterative process to compute the numerical parameters characterizing the algorithm's underlying model. The Machine-learning models developed so far can be broadly classified as either generative or discriminative. Generative methods which are exemplified by Naive Bayes classifiers and Hidden Markov Models (HMMs) involved the creation or generation of rich models of probability distributions. Discriminative methods including the Logistic regression and conditional random fields (CRFs) were involved in the direct estimation of posterior probabilities based on observations (Prakash et al, 2011).

Recent research has shifted the focus to unsupervised and semi-supervised learning algorithms which are able to learn from data that has not been hand-annotated with the desired answers, or using a combination of annotated and non-annotated data. Modern NLP algorithms which are based on machine learning, and specifically statistical machine learning, are different from the previously existing language processing tools which involved the direct hand coding of large sets of rules. The modern machine-learning algorithms which have been applied to NLP take a large set of features as input and have the advantage of expressing the relative certainty of different possibilities rather than only one thereby producing more reliable results on inclusion in a larger system.

The revival of MT could be attributed to the contributions of the Europeans and Japanese, particularly the European Commission which not only used customized pragmatism based production systems but also promoted the Eurotra research project on multi-lingual translation within a common, well-defined transfer framework. Many active Japanese teams also developed some translation products which were marketed (Nagao, 1989). The work on MT was based on the assumption that useful material related to specific application could be provided with or without user participation in the translation process and reflected the current state of NLP in grammar choices and the use of modular system architectures.

C. *The 1990s*

Although the early theories and methods of NLP were derived from linguistics, a major shift in its nature and processes was experienced in the early 1990s with an increasing inclination towards empirical methodologies in comparison to the introspective generalizations used in Chomsky's era. The focus in NLP shifted to performance data observed in naturally occurring text from the grammatically acceptable language and became more widely used with the advent of larger corpora.

D. *The 2000s*

The increased use of statistical methods which could be attributed to the availability of larger, performance-oriented corpora led to the establishment of these methods as very efficient and capable of accomplishing language analysis tasks at human performance level. Performance comparable to humans was shown to be achieved by the early statistical Part-Of-Speech tagging algorithms using Hidden Markov Models. Extensive electronic, lexical and syntactic resources led to the development of a state-of-the-art statistical parser which was seen to perform more accurately than a broad coverage rule-based parser on the test sections of the Penn TreeBank and also on unseen portions of the Brown Corpus (Ringger et al., 2004). Large corpora like Brown corpus and ongoing provision of collections funded by DARPA research programs contributed to the extensive electronic resources while WordNet provided the lexical-semantic knowledge bases which made the lexical resources. Gold standard syntactic resources were provided by the Penn TreeBank that contributed to the development of increasingly rich algorithmic analysis tools.

Another important milestone in the decade was the advent of the Bayes classifiers. These classifiers which are based on Bayes probability theorem can be trained to label an incoming text corpus. The classification is primarily done by calculating the probability that the text corpus belongs to each of the categories with which the system has been trained. The text corpus is then allocated to the category which has the highest probability. This algorithm marked an important breakthrough in the fields of disease prediction and document classification and made it possible to analyze emails and SMS and label them as spam.

The decade also saw the introduction of n-gram model. Using statistical properties of n-grams, this model helps in predicting the n^{th} text by analyzing the previous 'n-1' texts. Simplicity and scalability of these models saw them being widely used in speech and text recognition.

III. RECENT ADVANCEMENTS

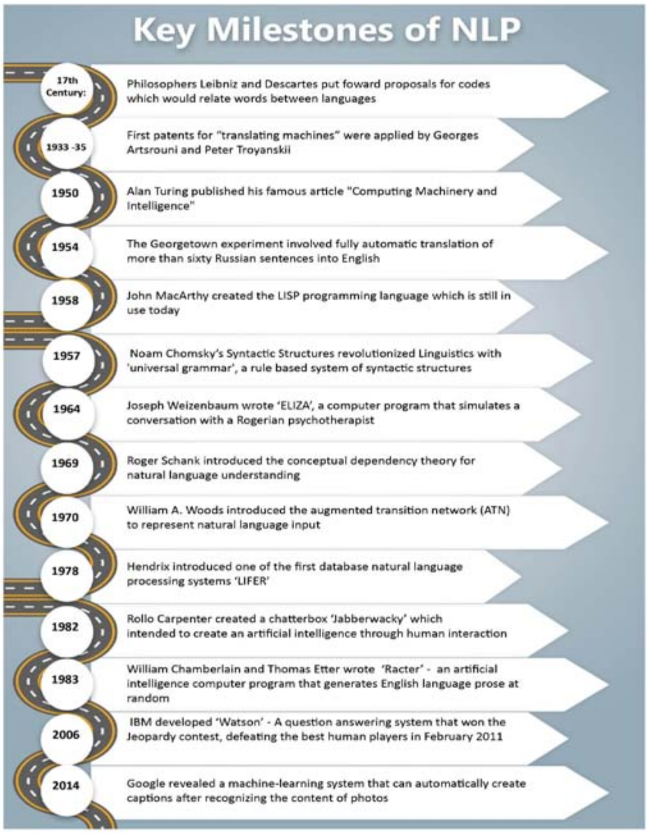
The last ten years of the millennium have seen immense growth in the field which can be attributed to an increased availability of large amounts of electronic text, high speed computers with larger memory and the advent of the Internet. The enormous improvements in machine technology have also allowed NLP to be performed with a true engineering spirit. The present advanced computing resources allow the running of powerful systems and conducting of impressively large experiments in a very short duration of time.

In the area of language processing, progress was made in the field of syntax with the development of effective grammar characterization means and techniques like chart parsing. Availability of extensive conceptual tools also contributed to the putting together of many systems or interface subsystems for experimental and development purposes. Statistical approaches succeeded in dealing with many generic problems in computational linguistics such as part-of-speech identification, word sense disambiguation, etc., and have become standard throughout NLP. NLP researchers are now developing next generation NLP systems that deal reasonably well with general text and account for a good portion of the variability and ambiguity of language. Speech understanding systems with language processing capabilities can now be thought of as the next level in this field.

Word2vec was another breakthrough in the field of NLP which has helped in surfacing semantic relationships by unsupervised processing of large amounts of text. This algorithm contains two distinct models (skip-gram and CBOW). These models have two different training methods (with/without negative sampling). Skip-gram model primarily predicts neighboring words around a word. In contrast, the CBOW model predicts the current word, given the neighboring word. Using simple vector difference, word2vec demonstrated that it is possible to get better word representations if a model's complexity is traded for efficiency. An adaptation of word2vec is doc2vec, an unsupervised algorithm which generates vectors for words. These vectors are then used for finding similarity between sentences. Doc2vec algorithm is an ideal choice when the input corpus has been proofread and has little or no spelling mistakes.

Another notable advancement has been the advent of approximate string matching, also known as ‘fuzzy string searching’. In this algorithm the accuracy of a match is determined by the number of primitive operations (also known as edit distance) required to convert a string into an exact match. This algorithm is widely used in spelling correction programs to find the best match for a text not found in dictionary.

Key milestones of NLP are depicted in Figure 1.



Key Milestones of NLP

17th Century: Philosophers Leibniz and Descartes put foward proposals for codes which would relate words between languages

1933 -35 First patents for "translating machines" were applied by Georges Artsrouni and Peter Troyanskii

1950 Alan Turing published his famous article "Computing Machinery and Intelligence"

1954 The Georgetown experiment involved fully automatic translation of more than sixty Russian sentences into English

1958 John MacArthy created the LISP programming language which is still in use today

1957 Noam Chomsky's Syntactic Structures revolutionized Linguistics with 'universal grammar', a rule based system of syntactic structures

1964 Joseph Weizenbaum wrote ‘ELIZA’, a computer program that simulates a conversation with a Rogerian psychotherapist

1969 Roger Schank introduced the conceptual dependency theory for natural language understanding

1970 William A. Woods introduced the augmented transition network (ATN) to represent natural language input

1978 Hendrix introduced one of the first database natural language processing systems ‘LIFER’

1982 Rollo Carpenter created a chatterbox ‘Jabberwacky’ which intended to create an artificial intelligence through human interaction

1983 William Chamberlain and Thomas Etter wrote ‘Racter’ - an artificial intelligence computer program that generates English language prose at random

2006 IBM developed ‘Watson’ - A question answering system that won the Jeopardy contest, defeating the best human players in February 2011

2014 Google revealed a machine-learning system that can automatically create captions after recognizing the content of photos

Figure 1. Key Milestones of NLP.

IV. Select Applications of NLP

Natural language processing can be applied to a wide range of processes which utilize text both theoretically and through implementation. The most commonly encountered applications of NLP include the following:

A. *Information Retrieval*

Since the 1940s, the need for automated information retrieval (IR) systems was felt which led to their development to aid in the management of huge scientific literature in that period. Their use has now extended to universities, corporate, and public libraries for provision of access to books, journals, and other documents. Commercial IR systems have now been developed which offer databases containing millions of documents in myriad subject areas by matching user queries with documents stored in a database. The algorithms used for IR can be widely classified into three classes.

1) Retrieval Algorithms: This is the main class of algorithms used in IR. They are involved in extraction of information from a textual database. These algorithms can be of two further types on the basis of extra memory required with them:

Sequential scanning of the text requires extra memory as a function of the query size rather than the size of the database. This can be exemplified by string searching with its running time being proportional to the text size.

Indexed text requires the availability of an "index" of the text which aids in speeding up the search. These are exemplified by inverted files and signature files with the index size being proportional to the database size and the search time being sub linear on the size of the text.

2) Filtering Algorithms: This class of algorithms works in a way that makes text, the input and a processed or filtered version of the text, the output. The purpose of these algorithms is size reduction of the text for simplification of the search. The most commonly used filtering/processing operations include the removal of common words using a list of stop words, transformation of uppercase letters to lowercase letters, removal of special symbols, reduction of sequences of multiple spaces to one space, transformation of numbers and dates to a standard format, transformation of spelling variants using Soundex-like methods, word stemming, automatic keyword extraction and word ranking.

However, there could be certain disadvantages to these filtering operations. For example, they require queries to be filtered too, like the text. Searching for common words, special symbols, or uppercase letters is also not possible through this method. Text fragments that have been mapped to the same internal form cannot be distinguished by this method.

3) Indexing Algorithms: These algorithms are involved in building data structures which aid in rapid text searching. On the basis of retrieval approaches, there could be many classes of indices like inverted files, signature files, keys, etc. These indices are mostly based on hashing or some kinds of tree.

This could become one of the greatest applications of NLP as it significantly involves text, but is still not utilized by many implementations. In recent times, significant systems based on NLP statistical approaches have been developed by Liddy (2001) and Strzalkowski.

B. *Information Extraction (IE)*

This area of application focuses on the recognition, tagging, and extraction of certain key information into a structured representation, which can then be utilized for a range of applications including question-answering, visualization, and data mining. It involves the extraction of structure from noisy, unstructured sources and has been researched upon extensively for over two decades due to its challenging nature. Emerging from the Natural Language Processing (NLP) community, IE is also engaged with other aspects of NLP including machine learning, information retrieval, database, web, and document analysis.

IE has evolved significantly in the past two decades to address the needs of these diverse applications. While the early systems involved manually coded rules, they were replaced by algorithms for automatically learning rules from examples due to their tedious nature. When rules became too brittle due to the targeting of more noisy unstructured sources by extraction systems, statistical learning emerged. This method deployed two kinds of techniques in parallel which were the Hidden Markov Models based generative models and 'maximum entropy' based conditional models. The global conditional models also known as Conditional Random Fields then superseded the previously existing methods. However none of these methods could be considered the final winner in the search for the best model for IE. These methods continue to be used in parallel in recent times with the development of hybrid models with benefits of both statistical and rule based methods depending on the nature of the extraction task (Sarawangi, 2007).

C. *Question Answering (QA)*

This computer science discipline focuses on building systems which can answer questions posed by humans in a natural language. Majority of the modern QA systems use natural language text documents as the underlying knowledge source.

NLP techniques are used to process questions, index or text corpus from which answers are extracted. An increasing number of QA systems use the World Wide Web as their corpus of text and knowledge.

A good search corpus is essential for QA. In the absence of documents containing the answer, there is very little that any QA system can do. Hence larger collection sizes generally result in better QA performance, unless the question domain is orthogonal to the collection. Due to data duplicity in massive collections, nuggets of information are phrased in various different ways in differing contexts and documents. This reduces the burden on the QA system to perform complex NLP techniques by making the correct information appear in various forms. This approach also helps in filtering correct answers from false positives.

In recent years, the scope of QA systems has been widened to encompass additional domains of knowledge. Systems that automatically answer geospatial, multilingual, biographical questions and questions about the content of video, audio and images have been developed.

D. *Automatic Summarization (AS)*

A part of data mining and machine learning, this computer program is used to create summaries by finding a subset of the data which succinctly encompasses the key information of the entire set. Length and syntax of texts along with the writing style are the key variables taken into account while creating a summary. AS finds its use in various applications such as search engines and news summary generators.

Extraction and Abstraction are the two prime approaches used in AS. In the former approach a subset of existing words -key phrases or sentences used in the original text- is selected to form the summary. In the latter approach an internal semantic representation is built which is then used by natural language generation techniques to create a summary. Summaries built using abstraction might contain words that are not present in the original document. Hence extractive summarization makes more sense in image collection summarization and video summarization.

E. *Machine Translation (MT)*

A sub-field of computational linguistics, MT is one of the oldest applications of NLP. The prime goal of MT systems is to provide the best possible translation from one language to another without any human assistance. MT systems primarily require a program for translation along with dictionaries and grammars to aid translation. MT systems are categorized into two categories – bilingual (translations between two specific languages) and multilingual (translations for any pair of language).

There are four types of translation approach used for MT-Direct Machine Learning Approach (oldest and less popular), Interlingua approach, Transfer approach and Empirical Machine Translation approach (emerging approach that uses large amount of raw data in the form of parallel corpora).

Current MT software often allows customization by domain or profession (e.g. weather report), thus improving the output by limiting the scope of allowable substitutions. Domains which use formulaic or formal language find this technique particularly effective. Besides being used by web translators, MT is now also being more readily used for translating government and legal documents.

V. Industrial Usage of NLP

NLP is primarily used to analyze text, thus allowing machines to understand human language. This human-machine interaction enables applications like sentiment analysis, parts-of-speech tagging, and topic extraction, named entity recognition, automatic text summarization, relationship extraction, stemming and more. NLP is commonly used for text mining, automated question answering and machine translation. Some of the well-known usages of NLP are:

- Conduct social media analysis through tools which track online brand conversations to understand what customers are saying and to glean insight
- Build RSS readers
- Build tools to improve quality of online communities by leveraging technology to 'auto filter' offensive conversations
- Build tool to track trending topics and popular hash tags
- Create chat bots
- Develop tools to monitor malicious digital attacks such as phishing or to detect whether somebody is lying or not
- Generate keyword tags
- Automated question answering bots such as Siri and Google Assistant

The applications and Industrial usage of NLP are summarized in Figure 2.

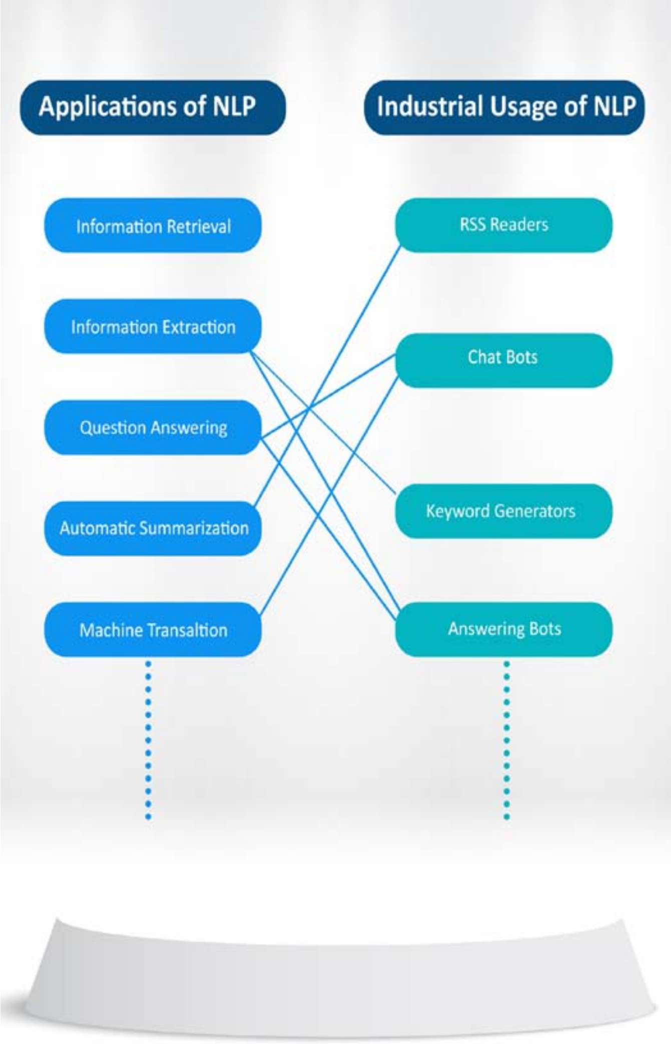


Figure 2. Applications and Industrial Usage of NLP.

VI. FUTURE PERSPECTIVE

The big step forward in Natural Language Processing will be devices that you talk to rather than type on. A tremendous amount of research is being conducted in the field of NLP for the development of new and improved systems which are more humanized and can understand simple instructions. Among the many notable research organizations working on different projects of NLP, the most remarkable is the Microsoft Natural Language Processing Group which aims to design and build a computer system that will analyze and generate Natural Languages. It is also working on broadening the scope of NLP through the development of parallel systems in languages like Chinese, English, French, German, Japanese, Korean and Spanish. Important research is also being carried out by the Canon Natural Language Processing Group which is focused on exploitation and further development of its language independent continuous speech recognition technology and interactive spoken systems.

Going ahead, commercialization of Natural Language Generation (NLG) will be another important development to keep an eye upon. NLG is the process of generating natural language from a machine representation system such as a knowledge base or a logical form. Till date the most successful application of NLG has been data-to-text system which performs text generation as well as data analysis by generating textual summaries of data sets and databases. One such application which uses NLG is UK Met Office's text-enhanced weather forecast.

In the coming years NLP will have a major impact on the Big Data economy. With advancement in related technologies such as Cognitive Computing and Deep Learning, NLP will provide a competitive advantage to businesses in the field of digital ad services, legal, media and medical science. Pricing trends can be predicted and advertising campaigns assessed by mining product reviews. It would become possible to predict candidate appeal and performance in elections by searching political forums. Social networks can be examined to find indicators of influence and power. Medical forums can be studied to discover common questions and misconceptions about sufferers of particular medical conditions so that website information can be improved. It would also be possible for computer systems to trade stocks and futures automatically, based on the sentiment of reports about companies.

VII. CONCLUSION

Despite its newness in comparison to other approaches of information technology, NLP can be said to have made significant progress and can boast of many successes. These advancements suggest the importance of NLP-based information access technologies and their relevance in the future as major areas of research and development in information systems.

VIII. REFERENCES

[1] A. M. Turing, Computing Machinery and Intelligence, Mind 49, pp. 433-460, 1950.

[2] N. Chomsky, Aspects of the theory of Syntax. Cambridge: M.I.T., 1965.

[3] N. Chomsky, Three models for the description of language, IRE Transactions on Information Theory, 2(3), pp. 113–124, 1956.

[4] C. Fillmore, The case for case, E. Bach and R. Harms (Ed.), Universals in linguistic theory, Holt, Rinehart, and Winston: New York, 1968.

[5] M. Kay, Functional Grammar, BLS 25, Linguistic Society of America, 1979.

[6] K. S. Jones, Natural Language Processing: a historical review, Current Issues in Computational Linguistics, 2001.

[7] W. C. Mann, S. Thompson, Rhetorical Structure Theory: Toward a Functional Theory of Text Organization, text 8(3), pp. 243-281, 1988.

[8] M. Nagao, (Ed). A Japanese view of machine translation in light of the considerations and recommendations reported by ALPAC, USA, Japan Electronic Industry Development Association (JEIDA), 1989.

[9] M. R. Quillian, Semantic memory, M. Minsky (Ed.), Semantic information processing, Cambridge, MA: MIT Press, 1968.

[10] P. M. Nadkarni, L. Ohno-Machado, W. W. Chapman, Natural language processing: an introduction, J Am Med Information Association 18(5): pp. 544–551, 2011.

[11] E. K. Ringger, R. C. Moore, E. Charniak, L. Vanderwende and H. Suzuki, Using the Penn Treebank to Evaluate Non-Treebank Parsers, Proceedings of Language Resources and Evaluation Conference (LREC), Lisbon, Portugal, 2004.

[12] R. C. Schank, Conceptual Information Processing, Amsterdam: North-Holland, print, 1984.

[13] S. Sarawangi, Foundations and Trends in Databases, vol 1, no 3, pp. 261–377, 2007.

[14] Y. A. Wilks, A preferential pattern seeking semantics for natural language inference, Artificial Intelligence 6(1), pp. 53-74, 1975.

[15] T. Winograd, Procedures as a Representation for Data in a Computer Program for Understanding Natural Language. MIT AITR-235, 1971.

[16] J. E. Weizenbaum, Communications of the Association for Computing Machinery, v.9 n.1, pp. 36-45, 1966.

[17] W. A. Woods, Transition Network Grammars for Natural Language Analysis. Communications of the ACM 13(10), 1970.

[18] J. Hirschberg, C. D. Manning, Advances in Natural Language Processing, 2015.