

Pattern Recognition and Natural Language Processing: State of the Art

Mirjana Kocaleva^{1,2}, Done Stojanov², Igor Stojanovik², Zoran Zdravev²

¹E-learning Center – University “Goce Delcev”, Krste Misirkov bb, Shtip, R.Macedonia

²Faculty of Computer Science – University “Goce Delcev”, Krste Misirkov bb, Shtip, R.Macedonia

Abstract – Development of information technologies is growing steadily. With the latest software technologies development and application of the methods of artificial intelligence and machine learning intelligence embeds in computers, the expectations are that in near future computers will be able to solve problems themselves like people do. Artificial intelligence emulates human behavior on computers. Rather than executing instructions one by one, as they are programmed, machine learning employs prior experience/data that is used in the process of system’s training. In this state of the art paper, common methods in AI, such as machine learning, pattern recognition and the natural language processing (NLP) are discussed. Also are given standard architecture of NLP processing system and the level that is needed for understanding NLP. Lastly the statistical NLP processing and multi-word expressions are described.

Keywords – artificial intelligence, machine learning, pattern recognition, natural language processing.

1. Introduction

Depending of the scope of application, there are many definitions for the artificial intelligence. According to [2], [4] artificial intelligence maps human behaviour on computers. Regardless whether the human behaviour is emulated or not, the goal of AI is to create intelligence. Without any doubt, the yet to come challenge in AI is to emulate completely or near-perfectly general intelligence. For different purposes, AI combines different methods from

linguistics, statistics and computational intelligence. AI is an interdisciplinary branch of computer science that has connections to other sciences such as neuroscience, philosophy, linguistics and psychology. Despite its application in industry, nowadays-predictive methods in AI are also commonly used in social sciences, such as economics.

There are several areas of specialization of artificial intelligence, such as:

- games playing, i.e. computers are programmed to oppose gamers
- Expert systems: computers are programmed to make decisions about situations in real-life (Mycin is a typical expert AI system that was developed in the 1970’s and it has been used for bacteria identification and recommendation of medications and drugs based on known symptoms).
- Natural language: computers are programmed to process sentences from spoken languages, analysing the morphology, lexicography and even the semantics of a whole sentence.
- Neural networks: combination of artificial neurons designed upon the neuron of a human being, primarily used for recognition purpose.
- Robotics: computers are programmed to receive surrounding signals and to generate intelligent reactions upon them.

Machine learning is a branch of AI that analyses systems, which are able to learn from training data, rather than following a strict order of execution of already programmed instructions. Machine learning tends to construct self-adjustable artificial systems, which can grow up alone and change through time by gathering experience from new training data sets [9]. Alike data mining, the machine learning also processes raw data in order to find patterns, but instead of extracting data which will undergo human analysis, it uses data to improve system’s own

DOI: 10.18421/TEM52-18

<https://dx.doi.org/10.18421/TEM52-18>

Corresponding author: Mirjana Kocaleva - E-learning Center – University “Goce Delcev”, Krste Misirkov bb, Shtip, R. Macedonia

Email: mirjana.kocaleva@ugd.edu.mk

 © 2016 Mirjana Kocaleva et al, published by UIKTEN. This work is licensed under the Creative Commons Attribution-NonCommercial-NoDerivs 3.0 License.

The article is published with Open Access at www.temjournal.com

capabilities. Detected patterns are used for system's self-adjustment [4], [5], [6].

There are two different kinds of learning. Learning without supervision and learning with adequate supervision. Learning without any kind of supervision requires an ability to identify patterns in streams of inputs, whereas learning with supervision involves classification and numerical regressions. The category an object belongs to is determined with classification and regression deals with obtaining a set of already recorded input/output samples, thereby from respective inputs we are discovering functions enabling the generation of suitable outputs.

Computational learning theory is a branch of modern computer science that deals with a mathematical analysis of the performance of machine learning algorithms. Essentially, machine learning can be defined as a self-learning data method for improving computer's actions or behaviour. The training data set strictly depends of the domain of the problem under consideration.

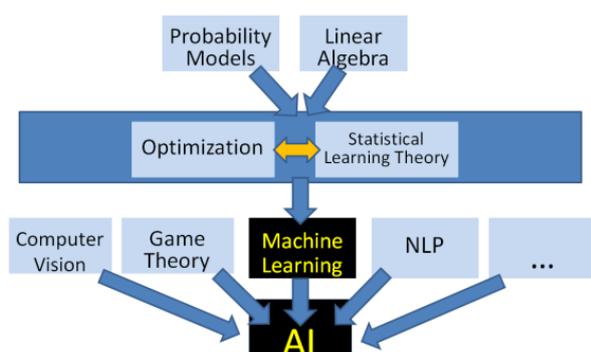


Figure 1. Retrieved from

<http://whiteswami.wordpress.com/machine-learning/>

Robot can learn to walk based on reading from sensors of force correction of the output for a specific input. Pattern recognition or the process of teaching a program or system to be able to recognize patterns [5], [6] is another way of thinking about machine learning.

2. Pattern recognition

Patterns are a form of language. Pattern recognition is studied in many fields, including psychology, psychiatry, ethnology, cognitive sciences, and computer science and traffic flow. Pattern recognition is a field in machine learning, but may also refer to pattern recognition (psychology), identification of faces, objects, words, melodies, etc. [8], [10]. Since the scopes of machine learning, knowledge discovery, pattern recognition and data mining highly overlap, they are hard to separate. Most often, machine learning refers to methods based on supervised learning, while unsupervised learning is primarily explored by data mining and KDD -

knowledge discovery. Unlike machine learning which is focused to maximize the rate of recognition [15], [17], [18], [19], pattern recognition models patterns and regularities found in data.

For our research, pattern recognition is important as a field in machine learning. Supervised learning employs training data set, which is used to identify patterns that match or resemble already annotated regularities. Unlike supervised learning, unsupervised learning does not rely on training data and it can be applied to detect unfamiliar regularities in data.

By analyzing training samples, supervised learning methods always produce an inferred function. Applying these functions, an output for any valid input object can be easily predicted. There are two kinds of inferred functions: if the output is discrete the function is called a classifier and if the output is continuous the function is called regression function. Data clustering (k-nearest neighbour's algorithm, support vector machine, naive Bayes classifier) and ANNs-artificial neural networks are common approaches to supervised learning.

According [2], unsupervised learning does not rely on previously labelled data and it attempts to detect built-in patterns in data. It can be used to calculate the correct output for any new data instance. Data clustering (k-means clustering and hierarchical clustering), hidden Markov models and blind signal separation using features extraction techniques for dimensionality reduction are common methods in unsupervised learning.

In pattern recognition, there are problems where distinct representations can be obtained for the same pattern, and depending on the type of classifier (statistical or structural), one type of representation is preferred versus the others. By taking into consideration the statistical variance of all inputs, algorithms for pattern recognition aim to perform "most probably" matching.

3. Natural language processing

Natural language processing (NLP) or computational linguistics is a major field of computer-related research. By discovering language patterns, children learn how to make a difference between singular and plural, match templates in nouns, verbs and adjectives and how to form different types of sentences, such as declarative or an imperative sentence. If we can define, and moreover, describe patterns from natural language then we can teach the computer about the way we speak and understand sentences form the spoken language. According [9], [13], [14] much of the research in this field relies on methods from cognitive science and

structure of this level is predictable and because of that reason, is used by NLP to understand the role of each information in a document [1], [3].

Pragmatic level

This level deals with the analysis of sentences and how they are used in different situations. In addition, also how their meaning changes depending on the situation [3].

All the levels described here are inseparable and complement each other. The aim of NLP systems is to inject these definitions into a computer, and then using them to create a structured unambiguous sentence with well-defined meaning [3], [7], [12], [16].

3.1. Learning about Natural Language Processing

As a discipline of informatics, NLP deals with the relationship among computers and natural languages. Understanding natural languages is an AI problem because the identification of natural languages requires understanding of sciences such as computer science, statistics, science of language and many others. Today's modern algorithms for NLP are based mainly on statistical machine learning.

Today's modern algorithms for NLP are mainly based on statistical machine learning. Machine learning aims to learn computer automatically from a large number of different corpus, in a different way from the methods of processing the natural language. A corpus is a set of documents that are annotated by hand with the correct values and it serves for learning purpose.

3.2. Statistical NLP

Statistical processing of natural language uses many different methods to solve problems. Some of those methods use probability, some use statistics, and others use mathematics and so on. Problems may arise with words that have multiple meanings and with sentences that are too long. Usually long sentences can be interpreted in several different ways. Methods for clarifying sentences usually use corpus and Markov models. Statistical NLP is composed of many techniques (such as modelling based on probability, algebra, and theory of information) for automated processing of natural or spoken language [7].

3.3. Multi-word expressions (MWE)

Multi-word expressions (MWEs) have paid attention from the NLP community. They are lexical elements and they:

- a) Can be degraded into several lexemes, and
- b) Can show lexical, syntactic, semantic, pragmatic and/or statistical idiomaticity [8].

When one or more parts of an MWE are not part of the conventional lexicon then we can say that lexical idiomaticity occurs. For example, in the word *Wi - Fi*, the components (*Wi* and *Fi*) do not have a meaning when they are separated.

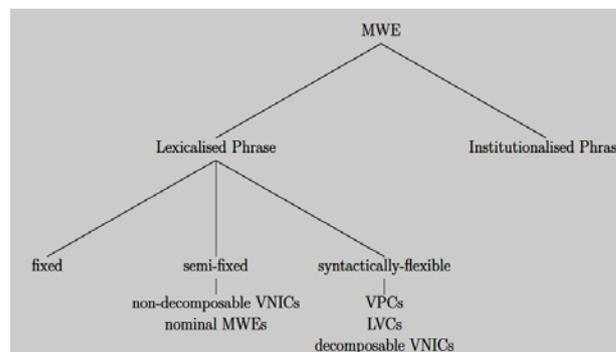


Figure 4. A classification for MWEs [8]

Syntactic idiomaticity happens when the syntax of the MWE can not be extracted from that of its building blocks [8], [16]. For example, we can analyse a syntactically idiomatic "on the whole". This word is adverbial but consists of a preposition (on) and an adjective (whole).

When the understanding of MWE is not coming from its building blocks, we are talking about semantic idiomaticity [8]. We can also use semantic idiomaticity for allegorical language [8].

When some MWE with a fixed location, position or situation are related, then we talk about pragmatic idiomaticity [8]. For example good night and good luck. The former is a greeting associated specifically with nights and the latter is having a good wishes for the people who work in mines.

Statistic idiomaticity occurs when some combination of words (term *under*) is repeated often in texts, unlike the words that make up the term *under*. Another example is *impeccable credentials*, which occur much more frequently than *spotless credentials*. Part of statistical idiomaticity are the binomials, for example true and false, where the opposite adjective does not keep the sense of the term *under* [8].

4. Conclusion

Study about natural language processing is progressively changing from semantics to narrative comprehension. In addition, people opinion is that the processing of a natural language represents a big AI difficulty. This is without any doubt a top challenge that yet to be solved- make computers smart and intelligent as people are. The future of NLP is

consequently link to the growth and evolution of AI. In addition, the future of MWE is nearly close to the development of NLP. For that reason, people have to understand at same time the information technologies (to know how computers work) and linguistic (as science of language).

Currently we think that there is still much work on this subject and with proper application of technologies for AI, such as pattern recognition, NLP and machine learning some concrete results may be obtained.

References

- [1]. S. Feldman, "NLP Meets the Jabberwocky: Natural Language Processing in Information Retrieval," *Information Today*, 1999.
- [2]. L. R. Ruiz, "Interactive Pattern Recognition applied to Natural Language Processing," Thesis, 2010.
- [3]. Rajesh.K.S, LokanathaC.Reddy, "Natural Language processing - an intelligent way to understand context sensitive languages," *International Journal of Intelligent Information Processing*, pp. 421-428, 2009.
- [4]. C. Janssen, "Artificial intelligence," 2014. [Online]. Available: <http://www.techopedia.com/definition/190/artificial-intelligence-ai>.
- [5]. C. Biemann, *Structure Discovery in Natural Language*, Springer, 2012.
- [6]. Alexander Clark, Chris Fox, Shalom Lappin, *The Handbook of Computational Linguistics and Natural Language Processing*, A John Wiley & Sons, Ltd., Publication, 2010.
- [7]. Joseph Olive, Caitlin Christianson, John McCary, *Handbook of Natural Language Processing and Machine Translation*, Springer, 2011.
- [8]. Chapman & Hall, *Handbook of natural language processing (second edition)*, CRC Press, 2010.
- [9]. James Pustejovsky and Amber Stubbs, *Natural Language Annotation for Machine Learning*, O'reilly, 2012.
- [10]. A. Kornai, *Mathematical Linguistics*, Springer, 2008.
- [11]. P. Pecina, *Lexical association measures Collocation Extraction*, 2009: Institute of Formal and Applied Linguistic.
- [12]. L. R. Ruiz, "Thesis - Interactive Pattern Recognition applied to Natural Language Processing," 2010.
- [13]. Steven Bird, Ewan Klein, and Edward Loper, *Natural Language Processing with Python*, O'reilly, 2009.
- [14]. V. Seretan, *Syntax-Based Collocation Extraction*, Springer, 2011.
- [15]. C. M. Bishop, *Pattern Recognition and Machine Learning*, Springer, 2006.
- [16]. P. M. Nugues, *An Introduction to Language Processing with Perl and Prolog*, Springer, 2006.
- [17]. Niels da Vitoria, Lobo TakisKasparis, Fabio Roli, James T. Kwok, Michael Georgiopoulos, Georgios C. Anagnostopoulos, Marco Loog, *Structural, Syntactic, and Statistical Pattern Recognition*, Springer, 2008.
- [18]. J. Tebelskis, "Speech Recognition using Neural Networks," Doctor Thesis, Pittsburgh, Pennsylvania, 1995.
- [19]. C. M. BISHOP, *Neural Networks for Pattern Recognition*, Oxford: Clarendon press, 1995.