The Title of Your Dissertation

Svetlozar Georgiev -
40203970

Submitted in partial fulfilment of
the requirements of Edinburgh Napier University
for the Degree of
BEng (Hons) Software Engineering

School of Computing

November 8, 2018

Authorship Declaration

I, Svetlozar Georgiev Georgiev, confirm that this dissertation and the work presented in it are my own achievement.

Where I have consulted the published work of others this is always clearly attributed;

Where I have quoted from the work of others the source is always given. With the exception of such quotations this dissertation is entirely my own work;

I have acknowledged all main sources of help;

If my research follows on from previous work or is part of a larger collaborative research project I have made clear exactly what was done by others and what I have contributed myself;

I have read and understand the penalties associated with Academic Misconduct.

I also confirm that I have obtained informed consent from all people I have involved in the work in this dissertation following the School's ethical guidelines.

Signed:

Date:

Matriculation no:

General Data Protection Regulation Declaration

Under the General Data Protection Regulation (GDPR) (EU) 2016/679, the University cannot disclose your grade to an unauthorised person. However, other students benefit from studying dissertations that have their grades attached.

Please sign your name below one of the options below to state your preference.

The University may make this dissertation, with indicative grade, available to others.

The University may make this dissertation available to others, but the grade may not be disclosed.

The University may not make this dissertation available to others.

Abstract

Contents

List of Tables

List of Figures

Acknowledgements

I would like to thank my cat, dog and family.

1 Introduction

2 Literature review

This chapter discusses the underlying techniques and methodologies...

2.1 Artificial intelligence

The Oxford dictionary [citation maybe?] defines *intelligence* as ”the ability to acquire and apply knowledge and skills”...

Artificial Intelligence (AI) is a multidisciplinary field whose goal is to automate activities that presently require human intelligence (Williams, 1983). (Poole et al., 1997) define AI as the study of the design of *intelligent agents* (or *rational agents*). Such agents receive percepts from the environment and perform actions, i.e. they implement a function that maps percept sequences to actions (Russel, Stuart and Norvig, Peter, 2003). The term *Artificial Intelligence* is also used to describe a property of machines or programs: the intelligence that the system demonstrates.

The term *artificial*, however, may introduce confusion as it suggests that it is not *real* intelligence. For this reason, different terms may be used in literature - *computational intelligence, synthetic intelligence,* etc.

(Williams, 1983) summarises the main concerns of AI as:

- Perception - building models of the physical world from sensory input.
- Manipulation - articulating appendages to affect desired state in the real world.
- Reasoning - understanding higher-level cognitive functions such as planning, drawing conclusions, diagnosing, etc.
- Communication - understanding and conveying information through the use of language.
- Learning - automatically improving a system's performance based on its experience.

The ultimate goal of AI is to understand the principles that make intelligent behaviour possible, in natural or artificial systems (Poole et al., 1997).

Several subfields of AI exist. The focus in this chapter will be on *Natural Language Processing (NLP)* and *Machine Learning* as they are relevant to the project.

2.2 Natural language processing

NLP is an area where AI, linguistics and Computer Science intersect. NLP focuses on making computers understand statements and words written in human language (Khurana et al., 2017). An NLP system should be able to determine the structure of text in order to answer questions about meaning or semantics of the written language (Martinez, 2010).

NLP consists of the following areas:

- *Speech recognition* - taking acoustic signal as input and determining what words were spoken.
- *Natural language understanding* - teaching computers how to understand natural (human language) instead of programming languages.
- *Natural language generation* - the process of producing phrases, sentences and paragraphs that are meaningful form an internal representation (Khurana et al., 2017)

Understanding human language is considered a difficult task due to its complexity. For example, words in a sentence can be arranged in an infinite number of ways. Additionally, words can have several meanings and contextual information is necessary to correctly interpret sentences.

The main techniques of understanding natural language are *Syntactic Analysis* and *Semantic Analysis*. The term *syntax* refers to refers to the grammatical structure of the text whereas the term *semantics* refers to the meaning that is conveyed by it. However, problems arise due to the fact that a sentence that is syntactically correct is not always semantically correct.

2.2.1 Syntactic analysis

Syntactic Analysis, also named Syntax Analysis or Parsing is the process of analysing natural language conforming to the rules of a formal grammar. Grammatical rules are applied to categories and groups of words, not individual words. Syntactic Analysis assigns a semantic structure to text.

2.2.2 Semantic Analysis

The word *semantic* is a linguistic term and means something related to meaning or logic. For humans, understanding what someone has said is an unconscious process that relies on intuition and knowledge about language itself. Therefore, understanding language is heavily based on meaning and context. Since computers can not rely on these techniques, they need a different approach.

Semantic Analysis can be defined as the process of understanding the meaning and interpretation of words, signs, and sentence structure. This enables computers partly to understand natural language the way humans do, involving meaning and context.

2.2.3 Techniques to understand text

- Parsing ...
- Stemming ...
- Text segmentation ...
- Relationship extraction ...

2.2.4 Natural Language Processing at the word and sentence level

The first step to implementing NLP is to parse the sentences into grammatical structures. However, parsing and understanding a natural language from an unbounded domain has proven extremely difficult as because of the complexity of natural languages, word ambiguity, and rules of grammar (Martinez, 2010).

Word sense disambiguation
Problems arise because words can have different meanings or senses based on the context, domain of discourse, etc. This is know as *polysemy*. The goal of *disambiguation* is to decide which of the meanings of a word should be attached to a specific use of the word. The methods to achieve can be categorised as *supervised*, *unsupervised* and *dictionary based*. Pantel and Lin approach...

Part-of-speech tagging
Some words can be used as a different part of speech...

Parsing
Parsing is the process of grouping sentence components into syntactic structures (Martinez, 2010). Some approaches to achieve this are *HMMs* and *probabilistic context-free grammars (PCFGs)*...

2.2.5 Natural Language Processing at the document level

...

2.3 Machine learning

Machine Learning is the science of getting computers to learn and act like humans do, and improve their learning over time in autonomous fashion, by feeding them data and information in the form of observations and real-world interactions [reference here]. Traditionally, algorithms are sets of explicitly programmed instructions used by computers to solve a specific problem. Machine learning algorithms, however, allow computers to be trained on data inputs and use statistical data analysis in order to output values which fall within a specific range. Because of this, machine learning facilitates computers in building models from sample data in order to automate decision-making processes based on data inputs.

The most widely adopted Machine Learning methods are:

- *Supervised learning* - trains algorithms based on example input and output data that is manually labelled by humans.
- *Unsupervised learning* - provides the algorithm with no labelled data in order to allow it to find structure within its input data.
- *Semi-supervised learning* -
- *Reinforcement learning* - the system attempts to maximize a reward based on its input data.

2.3.1 Supervised learning

In supervised learning, the computer is provided with example inputs that are labelled with their desired outputs. The purpose of this method is for the algorithm to be able to "learn" by comparing its actual output with the "taught" outputs to find errors, and modify the model accordingly. Supervised learning therefore uses patterns to predict label values on additional unlabelled data. A common use case of supervised learning is to use historical data to predict statistically likely future events.

Classification

The classification based tasks are a sub-field under supervised Machine Learning, where the key objective is to predict output labels or responses that are categorical in nature for input data based on what the model has learned in the training phase. Output labels here are also known as classes or class labels are these are categorical in nature meaning they are unordered and discrete values. Thus, each output response belongs to a specific discrete class or category (Kononenko and Kukar, 2007). Popular classification algorithms include *logistic regression, support vector machines, neural networks, ensembles* such as *random forests* and *gradient boosting, K-nearest neighbours, decision trees*, etc.

Regression

Machine Learning tasks where the main objective is value estimation can be termed as regression tasks. Regression based methods are trained on input data samples having output responses that are continuous numeric values unlike classification, where we have discrete categories or classes. Regression models make use of input data attributes or features (also called explanatory or independent variables) and their corresponding continuous numeric output values (also called as response, dependent, or outcome variable) to learn specific relationships and associations between the inputs and their corresponding outputs. With this knowledge, it can predict output responses for new, unseen data instances similar to classification but with continuous numeric outputs (Kononenko and Kukar, 2007).

- *Simple linear regression*
- *Multiple regression*
- *Polynomial regression*
- *Non-linear regression*
- *Lasso regression*
- *Ridge regression*
- *Generalised linear models*

2.3.2 Unsupervised learning

In unsupervised learning, data is unlabelled, so the learning algorithm is left to find commonalities among its input data. As unlabelled data are more abundant than labelled data, machine learning methods that facilitate unsupervised learning are particularly useful.

Without being given a specific "correct" answer, unsupervised learning methods can analyse complex data that is more expansive and seemingly unrelated in order to organise it in potentially meaningful ways. Unsupervised learning is often used for anomaly detection including for fraudulent credit card purchases, and systems that recommend what products to buy next.

The aim of supervised is extracting meaningful insights from data, rather than trying to predict an outcome based on training data (Kononenko and Kukar, 2007). More uncertainty exists with unsupervised training, however....

Unsupervised learning methods can be categorised as:

- *Clustering* -
- *Dimensionality reduction* -
- *Anomaly detection* -
- *Association rule-mining* -

2.3.3 Semi-supervised learning

Semi-supervised learning methods combine a lot of unlabelled data and a small amount of labelled and pre-annotated data. The possible techniques that can be used are *graph-based methods*, *generative methods*, and *heuristic-based methods*. A simple approach would be building a supervised model based on labelled data, which is limited, and then applying the same to large amounts of unlabelled data to get more labelled samples, train the model on them and repeat the process. Another approach would be to use unsupervised algorithms to cluster similar data samples, use human-in-the-loop efforts to manually annotate or label these groups, and then use a combination of this information in the future. This approach is used in many image tagging systems.

2.3.4 Reinforcement learning

Reinforcement Learning is a type of machine learning technique that enables an agent to learn in an interactive environment by trial and error using feedback from its own actions and experiences.

Typically the agent starts with a set of strategies or policies for interacting with the environment. On observing the environment, it takes a particular action based on a rule or policy and by observing the current state of the environment. Based on the action, the agent gets a reward, which could be beneficial or detrimental in the form of a penalty. It updates its current policies and strategies if needed. This iterative process continues until it learns enough about its environment to get the desired rewards.

As compared to unsupervised learning, reinforcement learning is different in terms of goals. While the goal in unsupervised learning is to find similarities and differences between data points, in reinforcement learning the goal is to find a suitable action model that would maximize the total cumulative reward of the agent.

2.4 Chatbots

...

2.4.1 History

3 Implementation

4 Results and discussions

5 Evaluation

6 Conclusions

7 Future work

References

Khurana, D., Koli, A., Khatter, K., and Singh, S. (2017). Natural Language Processing: State of The Art, Current Trends and Challenges. (Figure 1).

Kononenko, I. and Kukar, M. (2007). *Machine Learning Basics.*

Martinez, A. R. (2010). Natural language processing. *Wiley Interdisciplinary Reviews: Computational Statistics*, 2(3):352–357.

Poole, D., Mackworth, A., and Goebel, R. (1997). *Computational Intelligence: A Logical Approach.* Oxford University Press, Inc., New York, NY, USA.

Russel, Stuart and Norvig, Peter (2003). *Artificial Intelligence 3rd edition.*

Williams, C. (1983). A brief introduction to artificial intelligence. In *OCEANS'83, Proceedings*, pages 94–99. IEEE.

Appendices

A Project Overview

A.A Example sub appendices

...

B Second Formal Review Output

Insert a copy of the project review form you were given at the end of the review by the second marker

C Diary Sheets (or other project management evidence)

Insert diary sheets here together with any project management plan you have