EURA NOVA

FACULTIES OF SCIENCE AND ENGINEERING

Master Thesis Offers 2018

For the latest information, please visit http://research.euranova.eu.
26th January 2018

CONTENTS

EURA NOVA

INTRODUCTION

EURA NOVA is a Belgian company founded in September 2008. Our mission is simple: ``Being a technological incubator focusing on the pragmatic use of knowledge''. Our research and engineering activities are linked to technological directions and industrial opportunities.

Please visit our website http://euranova.eu for more information on our activities.

OUR MASTER THESIS OFFERS

This document presents Master Thesis offers supervised by our Research and Development department. Each of these projects represents a concrete opportunity to generate impact for EURA NOVA.

The student will work in close cooperation with the engineering team and will communicate the advance through the in-house EURA NOVA knowledge management tool.

For more information on our R&D activities, please visit our departement's website at http://research.euranova.eu.

HOW TO APPLY

If you are interested in one of our offers, please contact us at career@euranova.eu.

ARTIFICIAL INTELLIGENCE

CHARACTER LEVEL EMBEDDING FOR LOW LEVEL NLP TASKS

Context: Word embedding and document embeddings are techniques that gained popularity in recent years. They have been applied successfully in high level task in natural language understanding[MCCD13, VVR16] such as text classification, automatic answering. In practice, data scientists go very often through low level tasks to improve the performance of given nlp system. Low level task can be for example Part Of Speech tagging[SZ14] or name entity recognition[dSG15]. Recent research shows that building natural language understanding at character level can be very attractive option especially in the context of low level tasks. Character level embedding has many advantages. For instance, it deals with unknown work issue, since each word is considered as composition composition of letters. For language such as chinese where words are not separated by spaces character level system allow to avoid certain preprocessing steps such as word segmentation[CXL+15].

In this subject we will work on POS tagging from a character level point of view. We will explorer on one or more possible solution by evaluating and benchmarks. Then, we will evaluate the improvement of such method on high level task.

Business Opportunity: Building a POS tagger package can be very productive for our clients willing to improve their NLP systems such as text classification (email routing, sentiment analysis, etc.). In addition to that, a character level technology block can be very interesting in domain specific languages (medical, insurance, etc) where many strategy of transfer learning can be applied.

Contribution: The objective of this master thesis is :

- Analyze state of the art in character level models for natural language understanding.
- Design and implement a solution for POS tagging.
- Benchmark with other other statgries (work embedding, tfidf).
- Intergrate POS tagger output in a high level task (to be determined).

UBUNTU DIALOGUE CORPUS: APPLICATION TO QUESTION ANSWERING

Context: Question Answering is one of the most challenging application of artificial intelligence to natural language understanding [JM09]. Yet, question answering systems are useful in many industrial application from building chatbot systems, automatic email reply [KYR+16], improving customer relations, etc. Many systems have been proposed in the litterature from simple rule based systems, Information retrieval based systems and more recently neural based systems.

In this project, we restrict ourselves the to answer selection task. where the goal is to find the best answer that match a candidate questions in a domain specific context with technical terms. We will focus on Ubuntu Dialogue Corpus [LPSP15] and we will explore and benchmark different neural network based approaches. A side application of this project can be also in identifying similar questions in a forum.

Business opportunity: A typical application of such systems is in marketing and customer services. For example such algorithms can be useful to identify customer requests and/or to automate answers leading to an effective cost reduction.

Contribution: The objective of this master thesis is:

- Brief state of the art on question answering questions.
- Study state of the art on neural based system applied to automatic answering / sentence matching.
- Design and implement a solution.
- Improve the solution by doing several iterations.

APPLICATION OF IMAGE SEMENTIC SEGMENTATION TO IMPROVE COMPUTER VISION TASKS

Context: Recently, Deep Learning algorithms have achieved state of the art on several computer vision tasks such as image classification[HZRS16] or triggered other such as image captioning [LSV+14]. Yet, computer vision is considered a solved problem and current state of art is still far from solving scene understanding problem. Scene understanding is a long term objective which may be useful in applications such as robot vision or autonomous driving.

One of the promising techniques that can help towards better scene understanding is the image semantic segmentation [LSD15]. In Image semantic segmentation the goal is to recognize image context at pixel level as opposed image level as in traditional object classification. It is known to be a more challenging problem. Recent research tend to use variants of convolutional neural networks and encoder decoder architecture to solve this problem [FEF+17]. Some other explorer the usage of adversarial architectures [LCK16].

The goal of this project is to build a semantic segmentation system by testing one or more existing approaches from the state of the art. The second phase of the project is evaluate the improvement of other CV tasks by incorporation information from semantic segmentation techniques.

Business Opportunity: Such application can be very useful for solving problem related to autonomous driving or driving assistance application with our clients.

Contribution: The objective of this master thesis is:

- State of the art on semantic segmentation
- Design and implement a solution for image segmentation
- Test and benchmark on common datasets
- Report performance improvement (or not) by combining with existing visual search algorithms or object classification task.

MACHINE LEARNING & DATA SCIENCE

SEMI-SUPERVISED LEARNING WITH LARGE DATASETS

Context: Today collecting, storing and processing large amount of data is very common for a majority of organization. But these data comes very often unclean and unlabeled (from supervised machine learning point of view). This situation is restricting data scientists potential to create efficient predictive models on several use cases. To avoid this limitation we may think about semi-supervised approaches.

Research community have developed various way to perform semi supervised learning [OSZ06] such as: self-training approaches [Hea91, Hea91], co-training [DR06], clustering based methods, etc.

In this project our goal is to build a generic semi supervised framework and evaluating it of various type of data (structured data, text, etc.). We will restrict ourselves in clas

Business Opportunity: Eura Nova teams are very often faced with challenges of annotation scarcity on many projects with our clients. By exploring and testing several semi supervised learning, Eura Nova teams will be able to expand their possibilities on real case problems.

Contribution: The objective of this master thesis is:

- High level state of the art on semi supervised learning
- Design and implement an SSL framework for the given requirements
- Test and benchmark on common classification datasets
- Optional, report results by comparing different SSL approaches

HYPER PARAMETER TUNING IN HIGH DIMENSIONAL SPACES

Context: Today collecting, storing and processing large amount of data is very common for a majority of organization. But these data comes very often unclean and unlabeled (from supervised machine learning point of view). This situation is restricting data scientists potential to create efficient predictive models on several use cases. To avoid this limitation we may think about semi-supervised approaches.

Research community have developed various way to perform semi supervised learning [OSZ06] such as: self-training approaches [Hea91, Hea91], co-training [DR06], clustering based methods, etc.

In this project our goal is to build a generic semi supervised framework and evaluating it of various type of data (structured data, text, images). We will restrict ourselves in classification context and explore / benchmark one or more approaches from the state of the art.

Business Opportunity: Eura Nova teams are very often faced with challenges of annotation scarcity on many projects with our clients. By exploring and testing several semi supervised learning, Eura Nova teams will be able to expand their possibilities on real case problems.

Contribution: The objective of this master thesis is:

- Study state of the art in hyper parameter tuning
- Design and implement python package with one or more strategies
- Benchmark and compare different approach in the case of tuning neural networks.

REFERENCES

[CXL+15] Xinxiong Chen, Lei Xu, Zhiyuan Liu, Maosong Sun, and Huanbo Luan. Joint learning of character and word embeddings. In IJCAI International Joint Conference on Artificial Intelligence, volume 2015-January, pages 1236--1242, 2015.

[DR06] Luca Didaci and Fabio Roli. Using Co-training and Self-training in Semi-supervised Multiple Classifier Systems. pages 522--530. Springer, Berlin, Heidelberg, 2006.

[dSG15] Cicero Nogueira dos Santos and Victor Guimarães. Boosting Named Entity Recognition with Neural Character Embeddings. 2015.

[FEF+17] Damien Fourure, Rémi Emonet, Elisa Fromont, Damien Muselet, Alain Tremeau, and Christian Wolf. Residual Conv-Deconv Grid Network for Semantic Segmentation. jul 2017.

[Hea91] Marti A Hearst. Noun Homograph Disambiguation Using Local Context in Large Text. In Proceedings of the 7th Annual Conference of the University of Waterloo Centre for the New OED and Text Research, pages 1--15, 1991.

[HZRS16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep Residual Learning for Image Recognition. In 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 770--778, 2016.

[JM09] Daniel Jurafsky and James H Martin. Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition. Speech and Language Processing An Introduction to Natural Language Processing Computational Linguistics and Speech Recognition, 21:0--934, 2009.

[KYR+16] Anjuli Kannan, Peter Young, Vivek Ramavajjala, Karol Kurach, Sujith Ravi, Tobias Kaufmann, Andrew Tomkins, Balint Miklos, Greg Corrado, Laszlo Lukacs, and Marina Ganea. Smart Reply-Automated Response Suggestion for Email. SIGKDD 2016: Proceedings of the 22th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pages 955--964, 2016.

[LCK16] Pauline Luc, Camille Couprie, and Laboratoire Jean Kuntzmann. Semantic Segmentation using Adversarial Networks. Arxiv, 2016.

[LPSP15] Ryan Lowe, Nissan Pow, Iulian Serban, and Joelle Pineau. The Ubuntu Dialogue Corpus: A Large Dataset for Research in Unstructured Multi-Turn Dialogue Systems. jun 2015.

[LSD15] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, volume 07-12-June-2015, pages 3431--3440, 2015.

[LSV+14] Jonathan Long, Evan Shelhamer, Oriol Vinyals, Alexander Toshev, Samy Bengio, Dumitru Erhan, Karel Lenc, Andrea Vedaldi, Emily Denton, Soumith Chintala, Arthur Szlam, Rob Fergus, Philipp Fischer, H Philip, Caner Hazırbas, Patrick Van Der Smagt, Daniel Cremers, Thomas Brox, Fandong Meng, Zhengdong Lu, Zhaopeng Tu, Hang Li, Qun Liu, Vijay Mahadevan, and Student Member. Show and Tell: A Neural Image Caption Generator. arXiv, 32(1):1--10, 2014.

[MCCD13] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781, pages 1--12, 2013.

[OSZ06] Chapelle Olivier, Bernhard Schölkopf, and Alexander Zien. Semi-Supervised Learning, volume 1. 2006.

[SZ14] CD Santos and Bianca Zadrozny. Learning Character-level Representations for Part-of-Speech Tagging. Proceedings of the 31st International Conference on Machine Learning, ICML-14(2011):1818--1826, 2014.

[VVR16] Soroush Vosoughi, Prashanth Vijayaraghavan, and Deb Roy. Tweet2Vec: Learning Tweet Embeddings Using Character-level CNN-LSTM Encoder-Decoder. jul 2016.