EURA NOVA

FACULTIES OF SCIENCE AND ENGINEERING

Master Thesis Offers 2019-2020

For the latest information, please visit https://research.euranova.eu.
19th April 2019

CONTENTS

EURA NOVA

INTRODUCTION

EURA NOVA is a data-driven Belgian company founded in September 2008 and located in Brussels, Marseille, and Tunis. Our mission is simple: bring life to our customers' great ideas, by offering best-in-class services in data science, software engineering, and data architecture. To do so, we invest significantly in in-house expertise and state-of-the-art knowledge. In line with this course of action, we offer academic programs in collaboration with universities. These offers include bootcamps, master theses topics, research internships, and PhDs topics. See below for details.

OUR MASTER THESIS OFFERS

This document presents master theses topics supervised by our research & development department. Each project is an opportunity to be actively involved in the development of solutions to address tomorrow's challenges in ICTs and to implement them today.

The students will work in a dedicated **international** team of engineers **with diverse expertise in machine learning, graph theory, artificial intelligence, high performance computing, etc.**

They will keep EURA NOVA informed of the project advancement and share their ideas and challenges using the in-house knowledge management tool. We value continuous learning and teamwork. We love to have a good time together. For more information on our R&D activities, please visit our website at https://research.euranova.eu.

HOW TO APPLY

When you have gone through our master thesis offers, pick your favourite three. Draft a short text for each one, stating why you find it interesting and what you would do about it. Send us this statement, along with your CV at career@euranova.eu.

ARTIFICIAL INTELLIGENCE

CHARACTER LEVEL EMBEDDING FOR LOW LEVEL NLP TASKS

Context

Word embedding and document embedding are techniques that gained popularity in recent years. They have been applied successfully in high-level tasks in natural language understanding [MCCD13, VVR16] such as text classification and automatic answering. In practice, data scientists go very often through low-level tasks to improve the performance of given NLP systems. Low-level tasks are, 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 a very attractive option especially in low-level tasks. Character-level embedding has many advantages. For instance, it deals with unknown word issues, since each word is considered as a composition of letters. For languages such as Chinese where words are not separated by spaces, a character-level-based system makes it possible to avoid some preprocessing steps, such as word segmentation [CXL+15].

We will work on POS tagging from a character-level point of view. We will explore one or more possible solutions and we will benchmark the solutions. Then, we will evaluate how the solutions improve high-level NLP tasks.

Business Opportunity

Building a POS tagger package can be very productive for our clients who want to improve NLP systems such as text classification (email routing, sentiment analysis, etc.). In addition to that, a character-level NLP module can be very useful in domain-specific languages (medical, insurance, etc) where many strategies of transfer learning can be applied.

Contribution

The objectives of this master thesis are as follows:

- Analyse state-of-the-art character-level-based models for natural language understanding.
- Design and implement a solution for POS tagging.
- Benchmark with other strategies (word embedding, TF-IDF).
- Integrate POS tagger output in a higher level task (to be determined).

DEEP NEURAL NETWORK APPROACH TO ANSWER EXTRACTION

Context

Question answering is one of the most challenging applications of artificial intelligence to natural language understanding [JM09]. Yet, question answering systems are useful in many industrial use cases such as building chatbot systems, automatic email replies [KYR+16], customer relationship management applications, etc. Many systems have been proposed in the literature from simple rule-based systems, information-retrieval-based systems, to the recent neural-based systems. In this project, we focus on the answer extraction task where the goal is to find the best answer inside a document that matches a factoid question. We apply our models to several datasets such as the Stanford dataset [RJL18], the TREC dataset[1], and the legal document dataset COLIEE[2]. We will benchmark different neural-network-based approaches.

Business Opportunity

Typical applications of such systems can be found in marketing and customer service. For example, such algorithms can be used to identify customer requests or to automate answers. This leads to cost reduction.

Contribution

The objectives of this master thesis are as follows:

- Explore state-of-the-art question answering systems.
- Study state-of-the-art neural-based models applied to automatic answering.
- Design and implement a solution.
- Improve the solution by doing several iterations.

[1] https://trec.nist.gov/
[2] https://sites.ualberta.ca/~miyoung2/COLIEE2018/

VISUAL DOMAIN ADAPTATION: APPLICATION TO ROAD SIGN DETECTION

Context

Domain adaptation is currently one of the hot topics in artificial intelligence research. In DA, the goal is to learn a model from a data source distribution that will perform well on a different but related data target distribution [TC13]. Domain adaptation is not a new area; it has been used in many industrial applications. We can find in the literature various approaches to the DA problem, such as techniques based on covariance alignment [SS16] and techniques based on subspaces alignment [FHST13].

With the rise of artificial neural networks and their success in computer vision tasks, deep domain adaptation has emerged as a new learning technique to address the scarcity of labeled data. The main goal is to leverage deep networks to learn more transferable representations by embedding domain adaptation in the pipeline of deep learning. Recently, several techniques have been proposed, ranging from using convolutional neural networks [HZRS16] to generative adversarial networks [DOWL17].

In this work, we will focus on the application of domain adaptation to computer vision tasks (mainly classification and detection). We will first establish a clear view of the state-of-the-art methods and a clear distinction between domain adaptation techniques and transfer learning techniques. Then, we will implement and benchmark several approaches on toy datasets for experimentation. Finally, we will seek to apply successful techniques to road sign detection. Optionally, an original contribution may lead to a scientific contribution.

Business Opportunity

Such an application can be very useful for solving clients' problems related to autonomous driving or to driving-assistance applications. More generally, domain adaptation will improve products when data labels are difficult to obtain.

Contribution

The objectives of this master thesis are as follows.

- Study state of the art domain adaptation and transfer learning techniques.
- Design and implement one or several solutions.
- Test and benchmark the solutions with common datasets.
- Build a model to improve road sign identification.

TOPIC MODELING WITH (DEEP) NEURAL NETWORKS

Context

Topic Modeling is an established area of text mining focused on unsupervised text representation learning. Generative methods like Latent Dirichlet Allocation [BNJ03] have been successfully used for a variety of tasks. Recently, neural topic models gained popularity. The early methods use autoencoders [RS08], a Restricted Boltzmann Machine [HS09], autoregressive models [LL12] and Deep Boltzmann Machine [SSH13]. Latest models [CLL+15, LBC17] lean towards simple and elegant representation of topics and usage of Deep Learning structures such as Convolutional Neural Networks and Recurrent Neural Networks. However, the neural topic models existing today are not yet capable of controling the quality of the topics.

Business Opportunity

Neural Topic Models add flexibility of training and resource management to the text analysis, allowing processing of large amounts of unsupervised data. Using such models will facilitate the document representation, specifically in case of unsupervised datasets, and the neural nature of them will integrate seamlessly into existing architectures.

Contribution

The objectives of the thesis and its contributions are the following:

- cover the state of the art in neural topic models and their evaluation;
- design/implement a neural topic model;
- improve a neural topic model according to considerations of model quality [RBH15, OGCC15];
- test and benchmark the model.

The main contribution will lie in the new and/or improved model architecture and evaluation framework.

MACHINE LEARNING & DATA SCIENCE

SEMI-SUPERVISED LEARNING WITH LARGE DATASETS

Context

Today, collecting, storing, and processing a large amount of data is very common for a majority of organisations. Very often this data comes unclean and unlabeled (from a supervised machine learning point of view). This situation is restricting data scientists' potential to create efficient predictive models on several use cases. Semi-supervised approaches could be a solution.

The research community has developed various ways to perform semi-supervised learning techniques [OSZ06] such as: self-training approaches [Hea91], co-training [DR06], clustering based methods, etc.

In this project, our goal is to build a generic semi-supervised framework and to evaluate it on various types of data (structured data, text, etc.). We will restrict ourselves to a classification context and explore 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 techniques, EURA NOVA teams will be able to extend their new knowledge to real-case problems.

Contribution

The objectives of this master thesis are as follows.

- Study a high-level view of the state-of-the-art semi-supervised learning techniques.
- Design and implement a semi-supervised learning framework for the given requirements.
- Test and benchmark the framework with common classification datasets.
- Report results by comparing different semi-supervised learning approaches (optional).

HYPERPARAMETER TUNING IN HIGH DIMENSIONAL SPACES

Context

One of the most important phases in predictive modelling is hyperparameter tuning. In many cases, meticulous fine-tuning is very important as a 1% improvement on classification accuracy may cause a big return on investment (ROI). Hyperparameter tuning is a painful time-consuming operation. It is still easy to automate such a process, but simple techniques of automation can be inefficient with models with a high number of hyperparameters (neural networks, for instance).

There are several ways to tune hyperparameters, ranging from a simple brute force approach, through random search [BY12], to a more sophisticated Bayesian optimisation [SLA12]. However, these techniques tend to be limited in high dimensional spaces. Recently, many papers presented radical approaches that explore the architecture of neural networks, such as evolution-inspired techniques [MLM+17] or reinforcement-learning based techniques [ZL16].

The goal of this work is to build a generic hyperparameter tuning package that can be used by different machine learning algorithms, including deep neural networks. To do so, we will explore and benchmark one or more tuning techniques, using different classification datasets. In addition to that, we will test different strategies by combining or modifying existing approaches.

Business Opportunity

The outcome of this project is useful in the general case of hyperparameter tuning and specifically for optimising neural network architectures. Automating hyperparameter tuning, especially in the case of a deep neural network, can be very useful for existing EURA NOVA projects.

Contribution

The objectives of this master thesis are as follows.

- Study state-of-the-art hyperparameter tuning techniques.
- Design and implement a hyperparameter tuning package with one or more strategies, in Python.
- Benchmark and compare different approaches to tune neural networks.

PARAMETER TUNING OF GRAPH PROCESSING FRAMEWORKS

Context

Graphs are fundamental and widespread structures that provide an intuitive abstraction for the modeling and analysis of heterogeneous and highly interconnected data.

Large graphs have emerged in various business and scientific domains, such as modeling metabolic pathways in biology, and discovering communities and influencers in social networks. Big data application such as fraud detection, trends prediction, recommendation, routing and optimization, require graph frameworks to be efficiently modeled and solved.

A plethora of graph processing frameworks were developed to efficiently perform large scale, ad-hoc, and distributed computations over large graph data.

However, these frameworks have hundreds of parameters and thousands possible configurations. Manually choosing the right parameter values that optimize the performance of graph applications is therefore a complex and time-consuming task.

Contribution

In this thesis, you will study the model-based and black-box optimization techniques and apply them to graph processing. The aim of the project is to design a framework for automatic parameter tuning of graph processing frameworks.

DISTRIBUTED DATA PROCESSING

OPTIMISE THE DISTRIBUTED PROCESSING PIPELINE

Context

With big data becoming more widespread in today's business world, distributed frameworks such as Apache Spark, Apache Samza or Apache Flink tend to become quite common. These frameworks are often combined with other tools such as Apache Kafka for data ingestion or Hadoop Distributed File System (HDFS) for storage. Let us call the combination of these distributed processing frameworks and tools a processig pipeline. There are many parameters to tune in such a pipeline. Because of this, tuning takes long and requires expertise. We would like to automate the tuning process of the whole processing pipeline. This includes the tuning of each of the components themselves, but also the tuning of the pipeline itself by choosing its different components. The tuning of the pipeline should lead to the best possible performance.

Existing literature shows that there are several approaches to tune distributed processing frameworks parameters; they can be classified in two sets. The first set includes some performance prediction models to be used in the optimisation process [WXH16, CSG+18]. In this case, we need a set of historical runs of workloads, but no additional runs are needed in the optimisation process. The second one includes black-box optimisation processes [HLL+11, LZM+14]. The workload, here, is run multiple times in the optimisation process.

Business Opportunity

Being able to obtain the optimal configuration for any workload would be useful for any company that runs applications on distributed processing frameworks. It would allow them to set the right parameters to get the best out of the available technologies.

Contribution

The objectives of this master thesis are as follows.

- To get familiar with the state-of-the-art distributed processing frameworks performance prediction techniques.
- To implement three of these state-of-the-art techniques
- To design and implement a solution that improves the most recent techniques, for one specific framework.
- To transpose this solution to other frameworks, or to extend the solution to handle a distributed processing pipeline.

EXPLORE INTERESTING PATTERNS IN STREAMS

Context

Streaming context where users have already defined a workload that consists of multiple CEP queries written in the language we are developing now in order to discover some interesting situations or system malfunctions. Or only a set of raw streams where events are flowing into a data lake, and analysing these streams in order to discover interesting situations, by correlating the events of the different streams, is needed online and offline (similar to Exploratory OLAPs).

Contribution

The objective of this master thesis is to discover some interesting patterns for the user based on their workload, and then to evaluate the effectiveness of these recommended patterns (represented in our pattern algebra).

COMPLEX EVENT PROCESSING BENCHMARK

Context

Despite the fact that there are many frameworks nowadays that can provide complex event processing features, there is no standard mechanism to assess and evaluate these frameworks quantitatively. With the recent advancements, especially on the distributed processing part, there is an increasing need for a solid benchmark that provides a biggest possible set of metrics that can help to:

- Evaluate newly implemented systems;
- Discover the best fit for a specific use-case given a set of requirements and the available resources.

Contribution

The objective of this master thesis is to design and implement a benchmark that is able to assess CEP frameworks. Some **but not all** of the metrics that must be provided are: throughput, latency, memory consumption, correctness of results.

Implementation

Build a complete pipeline that generates data and route it into a black box (CEP framework to be assessed), in order to collect statistics and summaries, and then generate a set of KPIs

CREATE A USER-FRIENDLY EVENT QUERY LANGUAGE (EQL) INTERFACE

Context

The language we are developing now in the LEAD track has the objective of covering as many operators needed in the market as possible. Thus, the focus for the time being is centered around covering and improving the current state of the art by providing unambiguous algebra and a physical execution plan that is mapped to a full stream processing engine's operators. However, we are still missing an important aspect in the work that is user friendliness. Having a user friendly language/interface is of a great importance for our work to get to industry. The goal of a user-friendly interface is to provide a good user experience by ascertaining the following attributes:

- Simple. A user-friendly interface is not overly complex, but instead is straightforward, providing quick access to common features or commands;
- Clean. A good user interface is well-organized, making it easy to locate different tools and options;
- Intuitive. In order to be user-friendly, an interface must be make sense to the average user and should require minimal explanation for how to use it;
- Reliable. An unreliable product is not user-friendly, since it will cause undue frustration for the user. A user-friendly product is reliable and does not malfunction or crash.

Contribution

The objective of this master thesis is to improve the language grammar or in other words make it sweeter for human use so things can be expressed more clearly, more concisely, or in an alternative style that some may prefer. This is supposed to be achieved by spreading some syntactic sugar here and there, and introducing a big set of defined functions that would make users' lives easier. Additionally, design and implement a user-interface that will help the user express their needs easily and more concisely.

RETHINKING MICROSERVICES AND OTHER SERVICES ORCHESTRATION WITH DIGAZU IN REACTIVE STREAMING

Context

digazu[3] is a batch and real-time data supply chain developed by EURA NOVA. It is designed to fulfil customers' big data needs, such as real-time reporting, data integration, data science models industrialization… digazu is built onto cutting-edge technologies such as Kafka, Flink, Kubernetes, HDFS, ensuring reliability, efficiency and scalability in production environments. While digazu was not initially designed to orchestrate microservices, its architecture is in some points similar to the architecture used in reactive streaming patterns used to orchestrate microservices[4].

Business Opportunity

Many customers are moving gradually to microservices. Designing orchestration and implementation patterns with a tool like digazu would help customers accelerate their transformation and help them define good practices and standards.

Objectives

The objectives of this master thesis are the following:

- Study state of the art microservices orchestration patterns
- Analyse gaps and opportunities for such orchestrations using digazu
- Implement additional modules in digazu for microservices orchestration
- Define standards and good practices to orchestrate microservices using digazu

[3] https://digazu.eu/

[4] https://medium.com/capital-one-developers/microservices-when-to-react-vs-orchestrate-c6b18308a14c

IMPLEMENTING GDPR COMPLIANT SQL PROXY IN SQL STREAMING ON FLINK FOR DIGAZU

Context

digazu[5] is a batch and real-time data supply chain developed by EURA NOVA. It is designed to fulfil customers' big data needs, such as real-time reporting, data integration, data science models industrialization… digazu is built onto cutting-edge technologies such as Kafka, Flink, Kubernetes, HDFS, ensuring reliability, efficiency and scalability in production environments.

Business Opportunity

Integrating a GDPR compliant SQL proxy in SQL streaming on Flink in the transformation layer of digazu would help customers define processes around data that are GDPR compliant and help demonstrate compliance with GDPR requirements.

Objectives

The objectives of this master thesis are the following:

- Analyze how ABAC models can be used in the context of GDPR and metadata management
- Implement a GDPR compliant SQL proxy in SQL streaming on Flink in the transformation layer of digazu
- Define standards and good practices to define and maintain metadata for GDPR compliance

[5]https://digazu.eu/

REAL-TIME DATA CLEANING ON FLINK FOR DIGAZU

Context

digazu[6] is a batch and real-time data supply chain developed by EURA NOVA. It is designed to fulfil customers' big data needs, such as real-time reporting, data integration, data science models industrialization… digazu is built onto cutting-edge technologies such as Kafka, Flink, Kubernetes, HDFS, ensuring reliability, efficiency and scalability in production environments.

Business Opportunity

Data quality is an important factor to leverage data science models. Integrating real-time data cleaning capabilities in digazu would help customers get more insight out of their data.

Objectives

The objectives of this master thesis are the following:

- Analyze the state-of-the-art data cleaning techniques
- Assess techniques that can be executed in real time
- Implement a data cleaning module on Flink in the transformation layer of digazu

[6]https://digazu.eu/

PSEUDO-ANONYMISATION ON STREAMS FOR DIGAZU

Context

digazu[7] is a batch and real-time data supply chain developed by EURA NOVA. It is designed to fulfil customers' big data needs, such as real-time reporting, data integration, data science models industrialization… digazu is built onto cutting-edge technologies such as Kafka, Flink, Kubernetes, HDFS, ensuring reliability, efficiency and scalability in production environments.

Business Opportunity

Adding pseudo-anonymization capabilities to digazu would help customers be compliant with GDPR regulations and facilitate the data provisioning to data scientists.

Objectives

The objectives of this master thesis are the following:

- Analyse the state of the art pseudo-anonymisation techniques
- Assess how such techniques can be executed in real-time streaming
- Implement a pseudo-anonymization module in digazu

[7] https://digazu.eu/

BENCHMARKING OF THE DIGAZU TECHNOLOGIES

Context

digazu[8] is a batch and real-time data supply chain developed by EURA NOVA. It is designed to fulfil customers' big data needs, such as real-time reporting, data integration, data science models industrialization… digazu is built onto cutting-edge technologies such as Kafka, Flink, Kubernetes, HDFS, ensuring reliability, efficiency and scalability in production environments.

Business Opportunity

Benchmarking the technologies used in digazu would allow to better define the use cases that can be covered by the platform, as well as its limitations.

Objectives

The objective of this master thesis is to benchmark the technologies used in digazu, using different combinations of parameters such as:

- Volume of data
- Data throughput
- Number of topics
- Number of transformations
- Number of CPU's
- RAM
- …

[8]https://digazu.eu/

DATA ANALYTICS LAB INTEGRATION ON DIGAZU

Context

digazu[9] is a batch and real-time data supply chain developed by EURA NOVA. It is designed to fulfil customers' big data needs, such as real-time reporting, data integration, data science models industrialization… digazu is built onto cutting-edge technologies such as Kafka, Flink, Kubernetes, HDFS, ensuring reliability, efficiency and scalability in production environments.

Business Opportunity

Integrating a DAL (Data Analytics Lab) in digazu would help customers accelerate the development of data science models and the industrialization of such models.

Objectives

The objectives of this master thesis are the following:

- Define the requirements for the integration of a DAL in digazu, taking into account the end-to-end process of data models building and industrialization
- Implement a DAL module for digazu, facilitating the provisioning of data to data scientists and facilitating the industrialization of the models

[9] https://digazu.eu/

DATA GOVERNANCE

GDPR-COMPLIANT ACCESS POLICY MANAGEMENT

Context
GDPR is now active in EU and worldwide as soon as a company deals with EU citizen data. Among other things, the GDRP enforce the right to access a data according to a context, a purpose and a legal ground. For instance, a marketeer, can access the table customer profile for marketing purpose, because the user signed a contract. The legal ground can grouped in:

- Contracts
- Explicit opt in
- Legitimate interest
- Regulatory obligation

In a privacy by design how to enforce automatically this legal ground within the access security management ?

In this master thesis, we propose to investigate how a RBAC (Role based Access Control) or an ABAC (Attribute Based Access Control) can be extended in order to support this kind of requirements. The student will study the existing offer in terms of meta data management, security and access policy systems in order to propose a solution.

Contribution
The objectives of this master thesis are to:

- Study the state of the art in meta data management, RBAC and ABAC
- Study the meta data management in Hadoop
- Propose a formalism to represent the concept of legal ground
- Propose an approach to integrate the legal ground within an access security policy

The project will be supported by our legal expert and by an expert in Hadoop. This MT could be proposed as workshop or conference paper.

DATA GOVERNANCE

GDPR-COMPLIANT ACCESS POLICY MANAGEMENT

Context
GDPR is now active in EU and worldwide as soon as a company deals with EU citizen data. Among other things, the GDRP enforce the right to access a data according to a context, a purpose and a legal ground. For instance, a marketeer, can access the table customer profile for marketing purpose, because the user signed a contract. The legal ground can grouped in:

- Contracts
- Explicit opt in
- Legitimate interest
- Regulatory obligation

In a privacy by design how to enforce automatically this legal ground within the access security management ?

In this master thesis, we propose to investigate how a RBAC (Role based Access Control) or an ABAC (Attribute Based Access Control) can be extended in order to support this kind of requirements. The student will study the existing offer in terms of meta data management, security and access policy systems in order to propose a solution.

Contribution
The objectives of this master thesis are to:

- Study the state of the art in meta data management, RBAC and ABAC
- Study the meta data management in Hadoop
- Propose a formalism to represent the concept of legal ground
- Propose an approach to integrate the legal ground within an access security policy

The project will be supported by our legal expert and by an expert in Hadoop. This MT could be proposed as workshop or conference paper.

DISTRIBUTED STORAGE NEW RELATIONAL APPROACHES & NOSQL

Context
These last years, we have witnessed the emergence of a new category of distributed storages. They have the same features as traditional RDBMS while providing the distributed approach of the NoSQL DB. However, their architecture are still different and they offer different tradeoffs.

In this master thesis we propose to study (1) The distributed PostreSQL, (2) VoltDB, (3) NuoDB and to compare it to (1) Cassandra and (2) Druid.

The study will focus on the comparison of the underlying architecture, the performance tests and the discussions on the results.

Contribution
The objectives of this master thesis are to:

- To study the architecture of (1) the distributed PostreSQL, (2) VoltDB, (3) NuoDB
- To study the architecture of (1) Cassandra and (2) Druid
- To propose a benchmark for highlighting the differences
- To propose a test bench architecture and implementation
- To deploy the 5 analyzed framework and to apply the performance test
- To discuss the results and to interpret them

This MT could be proposed as workshop or conference paper.

REFERENCES

[BNJ03] David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. Journal of machine Learning research, 3(Jan):993--1022, 2003.

[BY12] James Bergstra JAMESBERGSTRA and Umontrealca Yoshua Bengio YOSHUABENGIO. Random Search for Hyper-Parameter Optimization. Journal of Machine Learning Research, 2012.

[CLL+15] Ziqiang Cao, Sujian Li, Yang Liu, Wenjie Li, and Heng Ji. A novel neural topic model and its supervised extension. In AAAI, pages 2210--2216, 2015.

[CSG+18] Zemin Chao, Shengfei Shi, Hong Gao, Jizhou Luo, and Hongzhi Wang. A gray-box performance model for apache spark. Future Generation Computer Systems, 2018.

[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.

[DOWL17] Alan Do-Omri, Dalei Wu, and Xiaohua Liu. A Self-Training Method for Semi-Supervised GANs. 2017.

[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.

[FHST13] Basura Fernando, Amaury Habrard, Marc Sebban, and Tinne Tuytelaars. Unsupervised visual domain adaptation using subspace alignment. In Proceedings of the IEEE International Conference on Computer Vision, 2013.

[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.

[HLL+11] Herodotos Herodotou, Harold Lim, Gang Luo, Nedyalko Borisov, Liang Dong, Fatma Bilgen Cetin, and Shivnath Babu. Starfish: a self-tuning system for big data analytics. In Cidr, volume 11, pages 261--272, 2011.

[HS09] Geoffrey E Hinton and Ruslan R Salakhutdinov. Replicated softmax: an undirected topic model. In Advances in neural information processing systems, pages 1607--1614, 2009.

[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.

[LBC17] Jey Han Lau, Timothy Baldwin, and Trevor Cohn. Topically driven neural language model. In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Association for Computational Linguistics, 2017.

[LL12] Hugo Larochelle and Stanislas Lauly. A neural autoregressive topic model. In Advances in Neural Information Processing Systems, pages 2708--2716, 2012.

[LZM+14] Min Li, Liangzhao Zeng, Shicong Meng, Jian Tan, Li Zhang, Ali R Butt, and Nicholas Fuller. Mronline: Mapreduce online performance tuning. In Proceedings of the 23rd international symposium on High-performance parallel and distributed computing, pages 165--176. ACM, 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.

[MLM+17] Risto Miikkulainen, Jason Liang, Elliot Meyerson, Aditya Rawal, Dan Fink, Olivier Francon, Bala Raju, Hormoz Shahrzad, Arshak Navruzyan, Nigel Duuy, and Babak Hodjat. Evolving Deep Neural Networks. arXiv, 2017.

[OGCC15] Derek O'Callaghan, Derek Greene, Joe Carthy, and Pádraig Cunningham. An analysis of the coherence of descriptors in topic modeling. Expert Systems with Applications, 42(13):5645--5657, aug 2015.

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

[RBH15] Michael Röder, Andreas Both, and Alexander Hinneburg. Exploring the space of topic coherence measures. In Proceedings of the Eighth ACM International Conference on Web Search and Data Mining - WSDM '15. ACM Press, 2015.

[RJL18] Pranav Rajpurkar, Robin Jia, and Percy Liang. Know What You Don't Know: Unanswerable Questions for SQuAD. arXiv, 2018.

[RS08] Marc' Aurelio Ranzato and Martin Szummer. Semi-supervised learning of compact document representations with deep networks. In Proceedings of the 25th international conference on Machine learning - ICML '08. ACM Press, 2008.

[SLA12] Jasper Snoek, Hugo Larochelle, and Ryan P Adams. Practical Bayesian Optimization of Machine Learning Algorithms. Adv. Neural Inf. Process. Syst. 25, pages 1--9, 2012.

[SS16] Baochen Sun and Kate Saenko. Deep CORAL: Correlation alignment for deep domain adaptation. In Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 2016.

[SSH13] Nitish Srivastava, Ruslan Salakhutdinov, and Geoffrey Hinton. Modeling documents with a deep boltzmann machine. In Uncertainty in Artificial Intelligence, page 616. Citeseer, 2013.

[SZ14] C D 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.

[TC13] Tatiana Tommasi and Barbara Caputo. Frustratingly easy NBNN domain adaptation. In Proceedings of the IEEE International Conference on Computer Vision, 2013.

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

[WXH16] Guolu Wang, Jungang Xu, and Ben He. A novel method for tuning configuration parameters of spark based on machine learning. In High Performance Computing and Communications; IEEE 14th International Conference on Smart City; IEEE 2nd International Conference on Data Science and Systems (HPCC/SmartCity/DSS), 2016 IEEE 18th International Conference on, pages 586--593. IEEE, 2016.

[ZL16] Barret Zoph and Quoc V Le. Neural Architecture Search with Reinforcement Learning. pages 1--16, 2016.