mathematics

Editorial

Preface to the Special Issue "Natural Language Processing (NLP) and Machine Learning (ML)—Theory and Applications"

Florentina Hristea [1,*] and Cornelia Caragea [2,*]

1 Department of Computer Science, Faculty of Mathematics and Computer Science, University of Bucharest, 010014 Bucharest, Romania
2 Department of Computer Science, University of Illinois at Chicago, Chicago, IL 60607, USA
* Correspondence: fhristea@fmi.unibuc.ro (F.H.); cornelia@uic.edu (C.C.)

Citation: Hristea, F.; Caragea, C. Preface to the Special Issue "Natural Language Processing (NLP) and Machine Learning (ML)—Theory and Applications". *Mathematics* **2022**, *10*, 2481. https://doi.org/10.3390/math10142481

Received: 12 July 2022
Accepted: 14 July 2022
Published: 16 July 2022

Natural language processing (NLP) is one of the most important technologies in use today, especially due to the large and growing amount of online text, which needs to be understood in order to fully ascertain its enormous value. During the last decade, machine learning techniques have led to higher accuracies involving many types of NLP applications. Although numerous machine learning models have been developed for NLP applications, recently, deep learning approaches have achieved remarkable results across many NLP tasks. This Special Issue has focused on the use and exploration of current advances in machine learning and deep learning for a great variety of NLP topics, belonging to a broad spectrum of research areas that are concerned with computational approaches to natural language.

The paper authored by Mothe [1] concentrates on better understanding information retrieval system effectiveness when taking into account the system and the query, while the other existing dimensions (document collection, effectiveness measures) are left in the background. The paper reviews the literature of the field from this perspective and provides a clear negative answer to the basic but essential question: "Can we design a transparent model in terms of its performance on a query?" The review concludes there is "lack of full understanding of system effectiveness according to the context although it has been possible to adapt the query processing to some contexts successfully". It equally concludes that, so far, neither the system component analysis, nor the query features analysis has proven successful "in explaining when and why a particular system fails on a particular query". This leaves room for further analyses, which prove to be necessary.

The paper authored by Donaj and Maučec [2] reports the results of a systematic analysis of adding morphological information into neural machine translation (NMT) system training, with special reference to languages with complex morphology. Experiments are performed on corpora of different sizes for the English–Slovene language pair, and conclusions are drawn for a domain-specific translation system and for a general-domain translation system. The authors conclude that NMT systems can benefit from additional morphological information when one of the languages in the translation pair is morphologically complex, with benefits depending on the size of the training corpora, on the form in which morphological information is injected into the corpora, as well as on the translation direction. We hope the conclusions of this paper will stimulate further research in order to see if they could apply to other language pairs containing English and highly inflected languages.

The paper authored by Nisioi et al. [3] studies the degree to which translated texts preserve linguistic features of dialectal varieties. The paper provides the first translation-related result (to the best of our knowledge), showing that translated texts depend not only on the source language, but also on the dialectal varieties of the source language, with machine translation being impacted by them. These authors show that automatically distinguishing between the dialectal varieties is possible, with high accuracy, even after translation. Another benefit of this research is represented by the release of a dataset of augmented annotations to the Proceedings of the European Parliament that cover dialectal speaker information.

The paper authored by Banbhrani et al. [4] proposes a novel sentiment classification approach, leading to a robust sentiment classification model for recommending courses with Taylor-chimp Optimization Algorithm enabled Random Multimodal Deep Learning (Taylor ChOA-based RMDL). Extensive experiments are conducted using the E-Khool dataset and the Coursera course dataset, with empirical results demonstrating that the proposed Taylor ChOA-based RMDL model significantly outperforms state-of-the-art methods for course recommendation tasks.

The paper authored by Haynes et al. [5] proposes an innovative classification pipeline for the automatic classification of national health service (NHS) feedback. This pipeline switches between different text pre-processing, scoring, and classification techniques during execution and, as a result, attains a high level of accuracy in the classification of an entire range of datasets. The paper equally analyzes the limiting factors that can intervene in the classification process, impacting its accuracy, and identifies them as being: (i) the imbalanced category distribution, (ii) the use of repeated common terms across different categories, and (iii) the subjective nature of the manual classification. This research ultimately provides NHS with a tool which is proven can be used in place of manual classification.

The paper authored by Dascalu and Hristea [6] discusses automatic Hate Speech detection, which has become a very important topic for major social media platforms and an increasingly hot topic for the field of NLP. These authors comment that recent literature on Hate Speech detection lacks a benchmarking system that can evaluate how different approaches compare against each other. Their study intends to determine if sentiment analysis datasets, which by far outweigh Hate Speech ones in terms of data availability, can help train a Hate Speech detection problem for which multi-task learning is used. The aim is ultimately that of establishing a Dataset list with different features, available for testing and comparisons, since a standardization of the test Datasets used in Hate Speech detection is still lacking, making it difficult to compare the existing models. The secondary objective of the paper is to propose an intelligent baseline as part of this standardization, by testing multi-task learning, and to determine how it compares with single-task learning with modern transformer-based models in the case of the Hate Speech detection problem.

The paper authored by Savini and Caragea [7] deals with automatic sarcasm detection, the results of which are equally scattered across datasets and studies. This paper brings in strong baselines for sarcasm detection based on BERT pre-trained language models. The discussed BERT models are further improved by fine-tuning them on related intermediate tasks before fine-tuning them on the target task. The employed technique relies on the correlation between sarcasm and (implied negative) sentiment and emotions. A transfer learning framework that uses sentiment classification and emotion detection as individual intermediate tasks, to infuse knowledge into the target task of sarcasm detection, is designed and explored. One of the main conclusions of this research is that, if the dataset size for the target task—sarcasm detection—is small, then intermediate task transfer learning (with sentiment as the intermediate task) can significantly improve the performance.

The paper authored by Škorić et al. [8] explores the effectiveness of parallel stylometric document embeddings in solving the authorship attribution task, by testing a novel approach on literary texts in 7 different languages. The authors conclude that the combination of word, lemma, and PoS-based document representations can model the language to a greater extent than any of them alone, especially with respect to the authorship attribution task. Most of the presented composition methods outperform the baselines, with or without mBERT inputs, which, surprisingly enough, are found to have no significant positive impact on the results of these methods. Another benefit of this research is represented by the creation of the multilingual document representations dataset (28,204 10,000-token documents), 133 literary document embeddings for 7 European languages and multilingually trained weights grouped by document representation and language, all of which can be used in future research in stylometry and in NLP, with focus on the authorship attribution task.

The paper authored by Badache et al. [9] refers to unsupervised and supervised methods to estimate temporal-aware contradictions in online course reviews. It studies contradictory opinions in MOOC comments with respect to specific aspects (e.g., speaker, quiz, slide), by exploiting ratings, sentiment, and course sessions where comments were generated. The contradiction estimation is based on review ratings and on sentiment polarity in the comments around specific aspects, such as "lecturer", "presentation", etc. The reviews are time dependent, since users may stop interacting and the course contents may evolve. Thus, the reviews taken into account must be considered as grouped by course sessions. The contribution is threefold: (a) defining the notion of subjective contradiction around aspects, then estimating its intensity as a function of sentiment polarity, ratings and temporality; (b) developing a data set to evaluate the contradiction intensity measure, which was annotated based on a user study; (c) comparing the unsupervised method with supervised methods with automatic criteria selection. The data set is collected from coursera.org and is in English. The results prove that the standard deviation of the ratings, the standard deviation of the polarities, and the number of reviews represent suitable features for estimating the contradiction intensity and for predicting the intensity classes.

The paper authored by Fuad and Al-Yahya [10] aims to explore the effectiveness of cross-lingual transfer learning in building an end-to-end Arabic task-oriented dialogue system (DS), using the mT5 transformer model. The Arabic-TOD dataset was used in training and testing the model. In order to address the problem of the small Arabic dialogue dataset, the authors present cross-lingual transfer learning using three different approaches: mSeq2Seq, Cross-lingual Pre-training (CPT), and Mixed-Language Pre-training (MLT). The conclusion of this research is that cross-lingual transfer learning can improve the system performance of Arabic in the case of small datasets. It is also shown that results can be improved by increasing the training dataset size. This research and corresponding results can be used as a baseline for future study aiming to build robust end-to-end Arabic task-oriented DS that refer to complex real-life scenarios.

The paper authored by Ouyang and Fu [11] is concerned with improving machine reading comprehension (MRC) by using multi-task learning and self-training. In order to meet the complex requirements of real-life scenarios, these authors construct a multi-task fusion training reading comprehension model based on the BERT pre-training model. The proposed model is designed for three specific tasks only, which leaves an open window toward future study. It uses the BERT pre-training model to obtain contextual representations, which are then shared by three downstream sub-modules for span extraction, yes/no question answering, and unanswerable questions. Since the created model requires large amounts of labeled training data, self-training is additionally used to generate pseudo-labeled training data, in order to improve the model's accuracy and generalization performance. An improvement of existing results occurs, in terms of the F1 metric.

The paper authored by Curiac et al. [12] discusses the evaluation of research trends by taking into account research publication latency. To our knowledge, this is the first work that explicitly considers research publication latency as a parameter in the trend evaluation process. A new trend detection methodology, which mixes auto-ARIMA prediction with Mann–Kendall trend evaluations, is presented. Research publication latency is introduced as a new parameter that needs to be considered when evaluating research trends from journal paper metadata, mainly within rapidly evolving scientific fields. The performed simulations use paper metadata collected from IEEE Transactions on Computer-Aided Design of Integrated Circuits and System and provide convincing results.

The paper authored by Masala et al. [13] introduces a method for discovering semantic links embedded within chat conversations using string kernels, word embeddings, and neural networks. The identification of these semantic links has become increasingly necessary since the mixture of multiple and often concurrent discussion threads leads to topic mixtures and makes it difficult to follow multi-participant conversation logs. The authors come to very clear conclusions: "string kernels are very effective at utterance level, while state-of-the-art semantic similarity models under-perform when used for utterance similarity. Besides higher accuracy, string kernels are also a lot faster and, if used in conjunction with a neural network on top of them, achieve state of the art results with a small number of parameters".

The paper authored by Vanetik and Litvak [14] uses deep ensemble learning in order to extract definitions from generic and mathematical domains. The paper concentrates on automatic detection of one-sentence definitions in mathematical and general texts, for which this problem can be viewed as a binary classification of sentences into definitions and non-definitions. Since the general definition domain and the mathematical domain are quite different, it is commented that transfer cross-domain learning performs significantly worse than traditional single-domain learning. The superiority of the ensemble approach for both domains is empirically shown, together with the fact that BERT does not perform well on this task. Experiments performed on four datasets clearly show the superiority of ensemble voting over multiple state-of-the-art methods.

Finally, the paper authored by Burdick et al. [15] presents a systematic analysis of different curriculum learning strategies and different batching strategies. The three considered tasks are text classification, sentence and phrase similarity, and part-of-speech tagging, for which multiple datasets are used in the experiments. The paper takes into account different combinations of curriculum learning and batching strategies across the three mentioned downstream tasks. While a single strategy does not perform equally well on all tasks, it is shown that, overall, cumulative batching performs better than basic batching. We especially retain the general conclusion that "the observed batching variation is something that researchers should consider" more in the future.

We hereby note the large range of research topics that have been touched within this Special Issue, showing the diversity and the dynamic of a permanently evolving field, which is giving one of the most important technologies in use today, that of natural language processing (NLP). This Special Issue has provided a platform for researchers to present their novel work in the domain of NLP and its applications, with a focus on applications of machine learning and deep learning in this field. We hope that this will help to foster future research in NLP and all related fields.

As Guest Editors of this Special Issue we would like to express our gratitude to the 47 authors who contributed their articles. We are equally grateful to a great number of dedicated reviewers, whose valuable comments and suggestions helped improve the quality of the submitted papers, as well as to the MDPI editorial staff, who helped greatly during the entire process of creating this Special Issue.

Funding: This research received no external funding.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Mothe, J. Analytics Methods to Understand Information Retrieval Effectiveness—A Survey. *Mathematics* **2022**, *10*, 2135. [CrossRef]
2. Donaj, G.; Sepesy Maučec, M. On the Use of Morpho-Syntactic Description Tags in Neural Machine Translation with Small and Large Training Corpora. *Mathematics* **2022**, *10*, 1608. [CrossRef]
3. Nisioi, S.; Uban, A.S.; Dinu, L.P. Identifying Source-Language Dialects in Translation. *Mathematics* **2022**, *10*, 1431. [CrossRef]
4. Banbhrani, S.K.; Xu, B.; Lin, H.; Sajnani, D.K. Taylor-ChOA: Taylor-Chimp Optimized Random Multimodal Deep Learning-Based Sentiment Classification Model for Course Recommendation. *Mathematics* **2022**, *10*, 1354. [CrossRef]
5. Haynes, C.; Palomino, M.A.; Stuart, L.; Viira, D.; Hannon, F.; Crossingham, G.; Tantam, K. Automatic Classification of National Health Service Feedback. *Mathematics* **2022**, *10*, 983. [CrossRef]
6. Dascălu, Ș.; Hristea, F. Towards a Benchmarking System for Comparing Automatic Hate Speech Detection with an Intelligent Baseline Proposal. *Mathematics* **2022**, *10*, 945. [CrossRef]
7. Savini, E.; Caragea, C. Intermediate-Task Transfer Learning with BERT for Sarcasm Detection. *Mathematics* **2022**, *10*, 844. [CrossRef]
8. Škorić, M.; Stanković, R.; Ikonić Nešić, M.; Byszuk, J.; Eder, M. Parallel Stylometric Document Embeddings with Deep Learning Based Language Models in Literary Authorship Attribution. *Mathematics* **2022**, *10*, 838. [CrossRef]

9. Badache, I.; Chifu, A.-G.; Fournier, S. Unsupervised and Supervised Methods to Estimate Temporal-Aware Contradictions in Online Course Reviews. *Mathematics* **2022**, *10*, 809. [CrossRef]
10. Fuad, A.; Al-Yahya, M. Cross-Lingual Transfer Learning for Arabic Task-Oriented Dialogue Systems Using Multilingual Transformer Model mT5. *Mathematics* **2022**, *10*, 746. [CrossRef]
11. Ouyang, J.; Fu, M. Improving Machine Reading Comprehension with Multi-Task Learning and Self-Training. *Mathematics* **2022**, *10*, 310. [CrossRef]
12. Curiac, C.-D.; Banias, O.; Micea, M. Evaluating Research Trends from Journal Paper Metadata, Considering the Research Publication Latency. *Mathematics* **2022**, *10*, 233. [CrossRef]
13. Masala, M.; Ruseti, S.; Rebedea, T.; Dascalu, M.; Gutu-Robu, G.; Trausan-Matu, S. Identifying the Structure of CSCL Conversations Using String Kernels. *Mathematics* **2021**, *9*, 3330. [CrossRef]
14. Vanetik, N.; Litvak, M. Definition Extraction from Generic and Mathematical Domains with Deep Ensemble Learning. *Mathematics* **2021**, *9*, 2502. [CrossRef]
15. Burdick, L.; Kummerfeld, J.K.; Mihalcea, R. To Batch or Not to Batch? Comparing Batching and Curriculum Learning Strategies across Tasks and Datasets. *Mathematics* **2021**, *9*, 2234. [CrossRef]