

Current Approaches and Applications in Natural Language Processing

Arturo Montejo-Ráez *  and Salud María Jiménez-Zafra * 

Departamento de Informática, SINAI Research Group, CEATIC, Universidad de Jaén, Campus Las Lagunillas s/n, 23071 Jaén, Spain

* Correspondence: amontejo@ujaen.es (A.M.-R.); sjzafra@ujaen.es (S.M.J.-Z.)

1. Introduction

Artificial Intelligence has gained a lot of popularity in recent years thanks to the advent of, mainly, Deep Learning techniques. These algorithms have broken many of the barriers in difficult computer based tasks such as computer vision, decision making or machine translation, among others. Nevertheless, many of the applications and problems overcome were already attempted with traditional algorithms in machine learning, heuristic approaches or knowledge-based systems. The big difference from previous approaches is that the current proposals are data-driven: they are able to learn from large amounts of data and build models to perform different tasks with a level of success never reached previously by other solutions.

This shift has been especially dramatic for Natural Language Processing (NLP). Linguistic-based methods have been surpassed by end-to-end architectures, where no prior knowledge on language is needed, although only when a massive amount of data is available. During the last two years we have witnessed the birth of amazing language models with impressive results in many different tasks, defining the new state-of-the-art in all of them. These models do not include, explicitly, traditional language processing tasks such as morpho-syntactic tokenization, lemmatization, stop-words removal, syntactic parsing, part of speech labeling, and other linguistic treatments on the text. New models seem to learn all of this linguistic information just from data.

Thus, NLP research has shown impressive improvements in many major tasks: machine translation, language modeling, text generation, sentiment/emotion analysis, natural language understanding, and question answering, among others. The advent of new methods and techniques such as graph-based approaches and reinforcement learning over deep learning architectures have boosted many of the tasks in NLP to reach human-level (and even further) performance. This has attracted the interest of many companies, so new products and solutions can profit from the advances of this relevant area within the artificial intelligence domain.

However, intensive research is still being conducted using deep learning approaches. Many new relevant features are being proposed, mainly related to stylometry, personality, or psicolinguistics. All of them are ad hoc features computed from texts that try to capture profile information, which, as we will see, can be used together with traditional machine learning algorithms to overcome user-centered tasks.

This Special Issue focuses on emerging techniques and trendy applications of NLP methods as an opportunity to report on all these achievements, establishing a useful reference for industry and researchers on cutting edge human language technologies. The contributions included in this issue propose new NLP algorithms and applications of current and novel NLP tasks. In addition, some trends, potential future research areas and new commercial products have been identified.



Citation: Montejo-Ráez, A.; Jiménez-Zafra, S.M. Current Approaches and Applications in Natural Language Processing. *Appl. Sci.* **2022**, *12*, 4859. <https://doi.org/10.3390/app12104859>

Received: 9 May 2022

Accepted: 10 May 2022

Published: 11 May 2022

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Copyright: © 2022 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

2. Review of Issue Contents

The contributions collected in this Special Issue tackle diverse tasks in NLP: text classification, text summarization, question and answering, machine translation, etc. We have organized these papers according to these topics.

2.1. Text Classification

Text classification is still a major concern in NLP research. Several contributions are related to this topic. For example, ref. [1] predict whether a patient had been diagnosed with a mental disorder and, if so, the specific mental disorder type. LIWC, spaCy, fastText, and RobBERT were used to analyze Dutch psychiatric interview transcriptions. LIWC, in combination with the random forest classification algorithm, performed the best in predicting whether a person had a mental disorder or not. SpaCy, in combination with random forest, best predicted which particular mental disorder a patient had been diagnosed with. When studying the results obtained with RobBERT and fastText, it was found, by applying LIME analysis, that the difference between mental disorder and no disorder was more prevalent in the manner of speaking than in the topics or the semantic content. Again, classical ML techniques such as Random Forest are still very useful.

Multimodal approaches are also present. In [2], a novel approach to fuse textual and visual features using a scaled dot-product attention mechanism is proposed. This is used in a multimodal classification system applied in fake news detection. The attention mechanism allows fine-grained combination of both visual embeddings and word embeddings taken from the image and text found in posts. The system achieves competitive results on the Weibo dataset.

Another paper studies the automatic detection of misogyny in web content by building an annotated corpus from several sources and then training a system for classifying texts [3]. The system is based on BERT embeddings and a final linear regression classifier. The results are good, although not comparable to other systems. A major contribution is the way in which the corpus is generated, which can allow for augmenting training datasets on misogyny detection.

Text classification can also be used to classify types of texts at high-level semantics. For instance, ref. [4] explores different machine learning and neural network techniques for the classification of strings as problems, non-problems, solutions, and non-solutions. The algorithm that provided the best results was a convolutional neural network.

To address the detection of fake news, the authors of [5] present a solution based on three steps: stance detection, author credibility verification, and machine learning classification. Stance detection verifies the relevance between the title and paragraphs of a news item; if there is a match, the next module checks whether the author is authentic to determine whether the news item should be believed or not. Finally, machine learning algorithms are used to classify the news item.

Text classification can also be applied to user profiling. A proposal for personality recognition relying on the dominance, influence, steadiness, and compliance (DISC) model together with a Bag-of-Words model of language is presented in [6]. Classical machine learning algorithms such as AdaBoost and Random Forests achieved good performance.

Topic detection is still stimulating research. Ref. [7] applies BERT word embeddings and a classical clustering algorithm (spherical k-means) to assign documents to topics. The proposal encodes documents as a linear combination of word embeddings and word frequencies in the document. Topics have been previously identified using the spherical k-means algorithm over all word embeddings in the corpus. Finally, documents are associated with topics using cosine distance. This method outperforms other approaches such as PLSA (Probabilistic Latent Semantic Analysis) and do not need to fine-tune the deep learning model.

In addition to systems and methods, this Special Issue includes some overviews. Related to this topic, ref. [8] provides a review on corpora related to deception detection on several approaches to the study of deception and on previous research into its

linguistic detection. Moreover, the author explores the linguistic cues of deception in the Spanish language.

One last contribution to text classification is the creation of a new multi-modal Wikimedia Commons dataset based on concrete/abstract words [9], along with a novel multi-modal pre-training approach based on curricular learning. The authors use the curricular learning method to train the model on the concepts through images and their corresponding captions to achieve multimodal language modeling. BERT and Resnet-152 models are employed in each modality and combined using attentional pooling to perform pre-training on the dataset.

2.2. Name Entity Recognition

Among major natural language understanding tasks, information extraction is still attracting much of the research. Named Entity Recognition (NER) is a central problem here. This Special Issue covers some novel approaches to NER in different languages. For instance, ref. [10] proposes an approach to entity linking (associate mentions in documents to existing entities in a knowledge graph) that profits from structural information of the graph, so correlation information between entities is enriched. No deep learning is used here, nor machine learning. It is a fully distance-based approach.

Nevertheless, transformers are the most prominent approach to NER. In [11], the task of a nested named entity recognition over two and four levels of annotation is accomplished by fine-tuning a BERT model. The results outperform state-of-the-art approaches such as Bi-LSTM-CRF. Thus, this approach is easier to generalize as it does not need specific feature extraction methods.

Another contribution to fine-grained NER is [12]. This work proposes a system for using character-level embeddings over LSTM networks multi-stacked for feature fusion. The unbalance problem usually found in fine-grained NER is solved by means of contextual information of coarse-grained named entities. The system is able to outperform other state-of-the-art NER systems.

To close the papers related to NER, an interesting overview is also included, but it is focused on the clinical domain [13]. The paper summarizes the current status of named entity recognition techniques and clinical relationship extraction in the clinical domain, discussing the existing models for the two tasks and their performances, the current challenges and future directions.

2.3. Question and Answering

Staying in natural language understanding tasks, Question and Answering (Q & A) systems still emerge as a continuous topic of research. In this regard, the paper by [14] proposes an attention model to solve question difficulty estimation in Question–Answering tasks. The method first relates question and information components using dual multi-head co-attention. Then, a self-attention model is applied over these relationships. This approach sets a new state-of-the-art in question difficulty estimation.

Expanding the number of question-answer pairs of Thai Question Answering corpora using Multilingual Text-to-Text Transfer Transformer (mT5) is the approach proposed by [15]. In addition, the authors propose a new syllable-level evaluation metric, which they consider more suitable for the Thai language because there is no ambiguity in syllable tokenization.

One last contribution to the Q & A topic is the paper by [16]. This paper introduces a privacy-preserving machine reading comprehension system capable of working with private data at a large scale and that is language independent.

2.4. Machine Translation

The problem of out-of-vocabulary (OOV) or rarely occurring words that limit the performance of neural machine translation models is known in automatic machine translation. The authors of [17] present a post-processing method for correcting machine translation re-

sults using a named entity recognition (NER) model to overcome this problem and conduct experiments on Chinese to Korean translation.

Another relevant issue is the estimation of the quality of a translation system. In [18], a pure performance comparison between several multilingual pretrained linguistic models (mPLM) is performed. As a result of the experiments, the authors confirm that the XLM-TLM model performs better and that the induced learning of cross-language alignment during pre-training had a positive impact. Furthermore, they perform experiments using mBART, and its additional noise schemes had a positive effect.

Bilingual embeddings are the subject of [19]. To train English–Welsh bilingual embeddings, the authors combine a Welsh corpus of approximately 145 million words with an English Wikipedia corpus. To learn the monolingual embeddings, they use word2vec and fastText. In addition, they explore three cross-language alignment strategies: cosine similarity, inverted softmax, and cross-domain similarity local scaling (CSLS). Different combinations of these approaches were evaluated on two tasks, bilingual dictionary induction and cross-lingual sentiment analysis. The best results were obtained using fastText monolingual embeddings and the CSLS metric.

2.5. Dialogue Systems

Conversational agents and chatbots are leveraging the research in dialogue systems. Two papers are included in this Special Issue with two totally different approaches. One is based on classical algorithms, and another uses a large language model. The former paper [20] proposes a system architecture for conversational agents that performs language understanding by intent detection and slot filling. The answering mechanism is based on a text retrieval engine (BM25). A classical CRF model is applied to perform the filling task, and the SVM algorithm was used for intent classification. No deep learning models were needed. The second approach is a novel task-oriented Arabic dialogue dataset (Arabic-TOD) and proposes an end-to-end generative dialogue system based on the multilingual mT5 [21]. The experiments show a performance comparable to high-resourced languages, such as English, and that a joint-training strategy with English and Chinese leads to better results.

2.6. Other Tasks

Explainability is a matter of study which is gaining deserved interest in recent years, in order to guarantee trustworthy systems. The work presented in [22] proposes a system to represent multilingual sentences using a natural machine language. The paper generates related universal concepts that are intuitive, according to human evaluation. Also related to explainability, the aim of the work presented in [23] is to provide people with an understandable representation of the complications of a disease. The authors present an approach to extract disease causal pathways, through cause–effect relation extraction, from documents on diabetes, kidney disease, heart disease, and arterial disease posted on Thai hospital web boards.

Related to Information Retrieval systems, we can find a proposal for query expansion [24]. This paper proposes a query expansion technique for Information Retrieval systems. A supervised expansion technique using the Naïve Bayes Multinomial Naïve Bayes algorithm is presented to extract relevant terms from the first documents retrieved by the initial query. In the evaluation of the proposed method, more accurate results are obtained compared to those achieved by the systems which participated in the TREC2017 Precision Medicine Track.

As an additional contribution, this time related to text summarization, is the work presented in [25]. It is a monolingual approach for abstractive summarization in Catalan and Spanish. The approach is based on a Transformer encoder–decoder pretrained and fine-tuned specifically for the language under studied. The performance of the monolingual models is compared with two of the most widely used multilingual models in text summarization, mBART, and mT5. Moreover, the authors present a new metric, content reordering, intended to help quantify the reordering of original content within an abstractive summary.

3. Conclusions

This Special Issue covers some of the most trending tasks in natural language processing: text classification, machine translation, information extraction, explainability, question and answering, or dialogue systems, among other topics. Many of the contributions in this Special Issue set a new state-of-the-art in targeted tasks.

It is remarkable how multilinguality is fostering research to cover what are considered “low-resourced” languages (i.e., those different from English or Chinese). In addition, we can confirm from the set of contributions that deep learning models (LSTM, BERT, mT5, among others) have irrupted the NLP arena to move approaches from computational linguistics to end-to-end solutions. Still, classical machine learning algorithms such as CRF, SVM, or Random Forest, just to cite few, are valid choices in many scenarios and are integrated in some competitive systems.

As a last remark, we find interesting the advent of hybrid approaches, such as those based in the combination of multiple features (word embeddings, char embeddings, BoW, etc.). In this ensemble of methods and techniques, graph-based and knowledge-based approaches deserve the focus of a growing number of studies.

As a main conclusion, this Special Issue offers a wide and varied insight into current NLP research, a domain of research which has already been considered as the main frontier in artificial intelligence.

Funding: This work was supported by Project LIVING-LANG (RTI2018-094653-B-C21) funded by MCIN/AEI/10.13039/501100011033 and by ERDF A way of making Europe, Fondo Social Europeo and Administration of the Junta de Andalucía (DOC_01073), Grant P20_00956 (PAIDI 2020) and grant 1380939 (FEDER Andalucía 2014-2020) from the Andalusian Regional Government.

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

References

1. Spruit, M.; Verkleij, S.; de Schepper, K.; Scheepers, F. Exploring Language Markers of Mental Health in Psychiatric Stories. *Appl. Sci.* **2022**, *12*, 2179. [\[CrossRef\]](#)
2. Wang, J.; Mao, H.; Li, H. FMFN: Fine-Grained Multimodal Fusion Networks for Fake News Detection. *Appl. Sci.* **2022**, *12*, 1093. [\[CrossRef\]](#)
3. Aldana-Bobadilla, E.; Molina-Villegas, A.; Montelongo-Padilla, Y.; Lopez-Arevalo, I.; Sordia, O.S. A Language Model for Misogyny Detection in Latin American Spanish Driven by Multisource Feature Extraction and Transformers. *Appl. Sci.* **2021**, *11*, 10467. [\[CrossRef\]](#)
4. Mishra, R.B.; Jiang, H. Classification of Problem and Solution Strings in Scientific Texts: Evaluation of the Effectiveness of Machine Learning Classifiers and Deep Neural Networks. *Appl. Sci.* **2021**, *11*, 9997. [\[CrossRef\]](#)
5. Islam, N.; Shaikh, A.; Qaiser, A.; Asiri, Y.; Almakdi, S.; Sulaiman, A.; Moazzam, V.; Babar, S.A. Ternion: An Autonomous Model for Fake News Detection. *Appl. Sci.* **2021**, *11*, 9292. [\[CrossRef\]](#)
6. Hernández, Y.; Martínez, A.; Estrada, H.; Ortiz, J.; Acevedo, C. Machine Learning Approach for Personality Recognition in Spanish Texts. *Appl. Sci.* **2022**, *12*, 2985. [\[CrossRef\]](#)
7. Cheng, Q.; Zhu, Y.; Song, J.; Zeng, H.; Wang, S.; Sun, K.; Zhang, J. Bert-Based Latent Semantic Analysis (Bert-LSA): A Case Study on Geospatial Data Technology and Application Trend Analysis. *Appl. Sci.* **2021**, *11*, 11897. [\[CrossRef\]](#)
8. Almela, Á. A Corpus-Based Study of Linguistic Deception in Spanish. *Appl. Sci.* **2021**, *11*, 8817. [\[CrossRef\]](#)
9. Sezerer, E.; Tekir, S. Incorporating Concreteness in Multi-Modal Language Models with Curriculum Learning. *Appl. Sci.* **2021**, *11*, 8241. [\[CrossRef\]](#)
10. Li, Q.; Li, F.; Li, S.; Li, X.; Liu, K.; Liu, Q.; Dong, P. Improving Entity Linking by Introducing Knowledge Graph Structure Information. *Appl. Sci.* **2022**, *12*, 2702. [\[CrossRef\]](#)
11. Agrawal, A.; Tripathi, S.; Vardhan, M.; Sihag, V.; Choudhary, G.; Dragoni, N. BERT-Based Transfer-Learning Approach for Nested Named-Entity Recognition Using Joint Labeling. *Appl. Sci.* **2022**, *12*, 976. [\[CrossRef\]](#)
12. Kim, H.; Kim, H. Fine-Grained Named Entity Recognition Using a Multi-Stacked Feature Fusion and Dual-Stacked Output in Korean. *Appl. Sci.* **2021**, *11*, 10795. [\[CrossRef\]](#)
13. Bose, P.; Srinivasan, S.; Sleeman, W.C.; Palta, J.; Kapoor, R.; Ghosh, P. A Survey on Recent Named Entity Recognition and Relationship Extraction Techniques on Clinical Texts. *Appl. Sci.* **2021**, *11*, 8319. [\[CrossRef\]](#)
14. Song, H.J.; Yoon, S.H.; Park, S.B. Question Difficulty Estimation Based on Attention Model for Question Answering. *Appl. Sci.* **2021**, *11*, 12023. [\[CrossRef\]](#)

15. Phakmongkol, P.; Vateekul, P. Enhance Text-to-Text Transfer Transformer with Generated Questions for Thai Question Answering. *Appl. Sci.* **2021**, *11*, 10267. [[CrossRef](#)]
16. Ait-Mlouk, A.; Alawadi, S.A.; Toor, S.; Hellander, A. FedQAS: Privacy-Aware Machine Reading Comprehension with Federated Learning. *Appl. Sci.* **2022**, *12*, 3130. [[CrossRef](#)]
17. Lee, J.; Lee, J.; Lee, M.; Jang, G.J. Named Entity Correction in Neural Machine Translation Using the Attention Alignment Map. *Appl. Sci.* **2021**, *11*, 7026. [[CrossRef](#)]
18. Eo, S.; Park, C.; Moon, H.; Seo, J.; Lim, H. Comparative Analysis of Current Approaches to Quality Estimation for Neural Machine Translation. *Appl. Sci.* **2021**, *11*, 6584. [[CrossRef](#)]
19. Espinosa-Anke, L.; Palmer, G.; Corcoran, P.; Filimonov, M.; Spasić, I.; Knight, D. English–Welsh Cross-Lingual Embeddings. *Appl. Sci.* **2021**, *11*, 6541. [[CrossRef](#)]
20. Chuang, H.M.; Cheng, D.W. Conversational AI over Military Scenarios Using Intent Detection and Response Generation. *Appl. Sci.* **2022**, *12*, 2494. [[CrossRef](#)]
21. Fuad, A.; Al-Yahya, M. AraConv: Developing an Arabic Task-Oriented Dialogue System Using Multi-Lingual Transformer Model mT5. *Appl. Sci.* **2022**, *12*, 1881. [[CrossRef](#)]
22. Qin, P.; Tan, W.; Guo, J.; Shen, B.; Tang, Q. Achieving Semantic Consistency for Multilingual Sentence Representation Using an Explainable Machine Natural Language Parser (MParser). *Appl. Sci.* **2021**, *11*, 11699. [[CrossRef](#)]
23. Pechsiri, C.; Piriyaikul, R. Causal Pathway Extraction from Web-Board Documents. *Appl. Sci.* **2021**, *11*, 10342. [[CrossRef](#)]
24. Silva, S.; Seara Vieira, A.; Celard, P.; Iglesias, E.L.; Borrajo, L. A Query Expansion Method Using Multinomial Naive Bayes. *Appl. Sci.* **2021**, *11*, 10284. [[CrossRef](#)]
25. Ahuir, V.; Hurtado, L.F.; González, J.Á.; Segarra, E. NASca and NAses: Two Monolingual Pre-Trained Models for Abstractive Summarization in Catalan and Spanish. *Appl. Sci.* **2021**, *11*, 9872. [[CrossRef](#)]