MDPI

Editorial

Preface to the Special Issue on Computational Linguistics and Natural Language Processing

Peter Z. Revesz

Department of Classics and Religious Studies, College of Arts and Sciences, and School of Computing, College of Engineering, University of Nebraska-Lincoln, Lincoln, NE 68588, USA; peter.revesz@unl.edu

Computational linguistics and natural language processing are at the heart of the AI revolution that is currently transforming our lives. We are witnessing the growth of these areas to the point that intelligent, talking robots can now perform many jobs that humans used to do. We are encountering robots in an increasing number of situations. For example, it is becoming common to see robots answer customer inquiries at call centers, replace cashiers with automated talking checkout machine at stores, look up a given address on an online map to plan a path and then autonomously navigate a car to its intended destination, assemble complex products while making decisions according to the particularities of supply and workflow demands in factories, and monitor access to buildings and sound an alarm if a dangerous situation develops. Robots have become such an integral part of our daily lives that while using the internet, people are required to take the "I am not a robot" test on a regular basis. This ambiguity resulting from the difficulty to distinguish humans and robots means that robots have acquired the capacity to replace humans.

In addition, intelligent and talking robots are increasingly used, even in situations and in places that are not as visible for most citizens. This includes various security systems, where robots are analyzing online conversations and chats. Some security robots also make decisions about potential dangers, such as possible illegal drug smuggling or acts of violence. Coupled with sophisticated drone systems, intelligent robots can assist humans by carrying out missions that require flying or being in outer space.

All of these machines need some way to communicate their contributions back to humans for the advancement of the human civilization. This is often best done by using human language, either written or spoken. Thus, the building of useful robots comes with the need to make the robot capable of using and analyzing human language. Therefore, the study of computational linguistics and natural language processing is a foundational part of the AI revolution that is presently resounding in our midst. It is in the light of these sentiments that this Special Issue on computational linguistics and natural language processing was called for, written, and assembled.

There is no doubt that computational linguistics and natural language processing facilitate not only major technological transformations but also influence social transformations. We increasingly live in what a few decades ago would have been termed a sci-fi world. These transformations come with certain challenges, but those challenges need not be feared because the opportunities outweigh the downside for humanity. That is also the overall message of the paper "Analyzing Sentiments Regarding ChatGPT Using Novel BERT: A Machine Learning Approach" by Sudheesh R. et al. [1] in this Special Issue, which is based on selected papers from the International Conference on Computational Linguistics and Natural Language Processing held in Beijing in December 2021. We also invited additional papers on the same topics, and they also underwent a rigorous review process.

Linguistic Profiling

Many of the papers proposed various novel methods of linguistic profiling and categorizing texts. The paper "Computing the Sound–Sense Harmony: A Case Study of William Shakespeare's Sonnets and Francis Webb's Most Popular Poems" by Delmonte [2]

proposes a novel sentiment analysis metric. The main idea is that sounds have presumed meanings, for example, they can be happy and sad. The actual text of a poem also expresses a meaning with a happy or sad connotation. Delmonte's sentiment analysis shows that both Shakespeare and Webb carefully chose their words to make the presumed meaning of the words' sounds and the sentences' explicit meaning be in harmony, or sometimes in disharmony when irony was intended.

The paper "Morphosyntactic Annotation in Literary Stylometry" by Gorman [3] provides a sophisticated stylometric analysis of texts. This computational analysis uses morphosyntactic annotations, such as the frequency of various pronouns, verbal cases, and the ordering of various elements of the sentence. Interestingly, famous authors are demonstrated to have a unique style because they can be identified by their stylometric profile with a very high probability. So, if we take a short quotation from one of their books, then they can be almost 100 percent correctly identified as the authors of their other books. For example, if we take a quotation from the book Oliver Twist by Charles Dickens, then we can identify that he also wrote A Christmas Carol because of the stylometric similarities between the two novels.

The above papers pose the prospect of intelligent robots imitating famous writers by simply adhering to some sentiment analysis measures and stylometric profiles as described in the two previous papers. The good news is that the paper "A Benchmark Dataset to Distinguish Human-Written and Machine-Generated Scientific Papers" by Abdalla et al. [4] proposes a practical approach to distinguish between human-written and machine-generated texts. While the paper focuses on identifying fake scientific papers, many of the proposed techniques can be applied to other types of texts too.

The paper "A Survey on Using Linguistic Markers for Diagnosing Neuropsychiatric Disorders with Artificial Intelligence" by Zaman and Trausan-Matu [5] is also concerned with the categorization of spoken language and written text. The goal of their paper is to aid medical diagnoses by identifying these linguistics markers, including sentiment analysis and stylometric measures, that characterize various mental illnesses. The paper also provides a comprehensive review of this growing subject with an extensive bibliography.

The paper "Linguistic Profiling of Text Genres: An Exploration of Fictional vs. Non-Fictional Texts" by Mendhakar [6] applies linguistic profiling to distinguish between texts that describe fiction versus those that describe real events. Various types of fiction are also categorized, such as fables, myths, mystery, romance, thriller, legends, and science fiction. Non-fiction works are also more finely divided into discussions, explanations, instructions, and persuasions.

The paper "A Literature Survey on Word Sense Disambiguation for the Hindi Language" by Gujjar et al. [7] focuses on the process to determine the exact context-specific meanings of ambiguous words, for example, the English word bark, which can mean the sound emitted by dogs, the outer sheath of a tree trunk, or a kind of ship. While many natural language processing techniques were developed to deal with the disambiguation of English words, the disambiguation of Hindi words sometimes requires language-specific algorithms that are reviewed in this paper.

Higher-Order Logical Representations and Methods

There were some papers that were concerned with the internal computer representation of texts and images. The paper "Agile Logical Semantics for Natural Languages" by Manca [8] introduces predicate abstraction as a new operator, which is argued to be a natural operator when some form of monadic high-order logic is used to express the semantics of linguistic structures. A possible application of predicate abstraction could be to teach more abstract logical thinking to chatbots, such as ChatGPT. For example, the author details a conversation with ChatGPT where ChatGPT was able to learn the predicate abstraction that "Go" and "Goes" represent the same predicate but in different grammatical forms.

Context-free Lindenmayer systems (D0L systems) have been used to describe the generation of growing plants, cities, and fractals, among other applications. Context-free Lindenmayer systems can be viewed as an extension of context-free grammars where the rewriting rules include commands such as 'draw a line', 'turn a specific degree', and 'go to a specific position'. They form a special type of language that is studied in the paper "D0L-System Inference from a Single Sequence with a Genetic Algorithm" by Łabędzki and Unold [9]. The aim of this paper is to essentially reverse engineer a context-free Lindenmayer system, that is, when given an image that was generated by a context-free Lindenmayer system, to then find its grammar. The authors demonstrate that their genetic algorithmic method finds satisfying solutions on various types of images, such as binary trees, Barnsley ferns, Koch snowflakes, and Sierpiński triangles.

Deciphering Scripts

A group of papers were concerned with the problem of deciphering inscriptions. The paper "Minoan Cryptanalysis: Computational Approaches to Deciphering Linear A and Assessing its Connections with Language Families from the Mediterranean and the Black Sea Areas" by Nepal and Perono Cacciafoco [10] provides a linguistic analysis of Linear A inscriptions, which were written by Minoan scribes during the Bronze Age, mainly on the island of Crete. The linguistic analysis uses the feature-based comparison of signs method introduced in Revesz [11]. However, in their study, Nepal and Perono Cacciafoco [9] obtain slightly different sign matches between Linear A signs, Carian alphabet letters, and Cypriot syllabary signs than Revesz [11] obtained, because they use different weights for the various features. The matches provide candidate phonetic values for Linear A signs. This allows for a phonetic transcription of Linear A inscriptions to be carried out.

Next, they applied a linguistic analysis of the Linear A inscriptions by finding possible words from the following languages: Ancient Egyptian, Hittite, Luwian, Proto-Celtic, and Proto-Uralic. The latter two languages were chosen because they may have been spoken on the coastal areas of the Black Sea, which has been shown to be the likely source of some Minoans [12,13]. The analysis yielded eight Ancient Egyptian, nine Hittite, seven Luwian, eleven Proto-Celtic, and twelve Proto-Uralic words as good matches with the Linear A inscriptions. While the analysis of Nepal and Perono Cacciafoco [10] is inconclusive in deciding the underlying language of the Linear A inscriptions, it nicely demonstrates that it was premature of many earlier authors to focus their attention only on Mediterranean languages, ignoring the fact that the Bosporus and Dardanelles Straits enable easy sailing between the Aegean Sea and the Black Sea areas. The analysis of Nepal and Perono Cacciafoco [10] is compatible with Revesz [11], which provided a translation of twenty-eight Linear A inscriptions into a Uralic language.

The paper "A Proposed Translation of an Altai Mountain Inscription Presumed to Be from the 7th Century BC" by Revesz and Varga [14] originated when someone brought to our attention an inscription from a book by Karžaubaj Sartkožauly, who is a member of the Kazakhstan Academy of Sciences. Sartkožauly presumed the inscription to be from the 7th century BC, and this was also our initial assumption. However, we became increasingly suspicious about the dating of the inscription during our decipherment work. For example, the inscription used a personal woman's name, Enikő, that was only created in the 19th century, although it became popular afterwards. This paper also proposed two different solutions because there was an ambiguous part in the inscription.

After the study by [14] was published, it received great publicity and became the subject of a popular YouTube video, which happened to be watched by the scribe, Peter Kun, who admitted in a comment below the video that he wrote the inscription while he was visiting the Altai Mountains as a young man. That was a fascinating turn of events because there is no other known case when a scribe practically "came alive" to be able to judge the correctness of the decipherment of a presumably ancient inscription.

The paper "Decipherment Challenges due to Tamga and Letter Mix-Ups in an Old Hungarian Runic Inscription from the Altai Mountains" [15] came as a natural follow up

of [14] after contacting Peter Kun. He provided a fascinating explanation of the cryptic section of his inscription, while verifying the correctness of the rest of our decipherment in [14]. I thought that the readers deserve to learn about the entire correct decipherment. In addition, ref. [15] also provides a mathematical analysis of the process of incorrectly mixing up some visually similar signs by Peter Kun. Since such mix-ups are frequent in other inscriptions too, this mathematical analysis may benefit other scholars who are working on the decipherment of ancient scripts.

Finally, I would like to thank the many reviewers who have reviewed the papers submitted to this Special Issue. I also would like to thank Janessy Zhan, Section Managing Editor at MDPI of this Special Issue, for her outstanding help in every aspect of organization, including arranging independent reviewers of my contributions to this Special Issue. I am also grateful to all parties that made contributions to this Special Issue. It was great to work with such a talented group of authors, and I wish them much success in their future research.

Conflicts of Interest: The author declares no conflict of interest.

References

1. Sudheesh, R.; Mujahid, M.; Rustam, F.; Shafique, R.; Chunduri, V.; Villar, M.G.; Ballester, J.B.; de la Torre Diez, I.; Ashraf, I. Analyzing Sentiments Regarding ChatGPT Using Novel BERT: A Machine Learning Approach. *Information* **2023**, *14*, 474.
2. Delmonte, R. Computing the Sound–Sense Harmony: A Case Study of William Shakespeare's Sonnets and Francis Webb's Most Popular Poems. *Information* **2023**, *14*, 576. [CrossRef]
3. Gorman, R. Morphosyntactic Annotation in Literary Stylometry. *Information* **2024**, *15*, 211. [CrossRef]
4. Abdalla, M.H.I.; Malberg, S.; Dementieva, D.; Mosca, E.; Groh, G. A Benchmark Dataset to Distinguish Human-Written and Machine-Generated Scientific Papers. *Information* **2023**, *14*, 522. [CrossRef]
5. Zaman, I.-R.; Trausan-Matu, S. A Survey on Using Linguistic Markers for Diagnosing Neuropsychiatric Disorders with Artificial Intelligence. *Information* **2024**, *15*, 123. [CrossRef]
6. Mendhakar, A. Linguistic Profiling of Text Genres: An Exploration of Fictional vs. Non-Fictional Texts. *Information* **2023**, *13*, 357. [CrossRef]
7. Gujjar, V.; Mago, N.; Kumari, R.; Patel, S.; Chintalapudi, N.; Battineni, G. A Literature Survey on Word Sense Disambiguation for the Hindi Language. *Information* **2022**, *14*, 495. [CrossRef]
8. Manca, V. Agile Logical Semantics for Natural Languages. *Information* **2024**, *15*, 64. [CrossRef]
9. Łabędzki, M.; Unold, O. D0L-System Inference from a Single Sequence with a Genetic Algorithm. *Information* **2023**, *14*, 343. [CrossRef]
10. Nepal, A.; Perono Cacciafoco, F. Minoan Cryptanalysis: Computational Approaches to Deciphering Linear A and Assessing its Connections with Language Families from the Mediterranean and the Black Sea Areas. *Information* **2024**, *15*, 73. [CrossRef]
11. Revesz, P.Z. Establishing the West-Ugric Language Family with Minoan, Hattic and Hungarian by a Decipherment of Linear A. *WSEAS Trans. Inf. Sci. Appl.* **2017**, *14*, 306–335.
12. Revesz, P.Z. Data Mining Autosomal Archaeogenetic Data to Determine Minoan Origins. In Proceedings of the 25th International Database Engineering & Applications Symposium, New York, NY, USA, 7 September 2021; Association for Computing Machinery: New York, NY, USA, 2021; pp. 46–55.
13. Revesz, P.Z. Minoan archaeogenetic data mining reveals Danube Basin and western Black Sea littoral origin. *Int. J. Biol. Biomed. Eng.* **2019**, *13*, 108–120.
14. Revesz, P.Z.; Varga, G. A Proposed Translation of an Altai Mountain Inscription Presumed to be from the 7th Century BC. *Information* **2022**, *13*, 243. [CrossRef]
15. Revesz, P.Z. Decipherment Challenges due to Tamga and Letter Mix-Ups in an Old Hungarian Runic Inscription from the Altai Mountains. *Information* **2023**, *13*, 422. [CrossRef]