ISSN 2007-9737

Computational Linguistics: Introduction to the Thematic Issue

This special issue of *Computación y Sistemas* presents a selection of latest papers on various areas of natural language processing and computational linguistics, along with two regular papers.

Natural language processing is an area of artificial intelligence and its applications devoted to analysis and generation of data streams involved in human communication using language, such as English or Spanish, typically in the form of text or speech, as well as, in multimodal setting, associated facial expressions and body language. It embraces natural language processing and computational linguistics in a wider context of their real-life applications in a wide range of disciplines.

The typical tasks of natural language processing include machine translation, text classification, text summarization, information extraction, author profiling, and sentiment analysis, while typical applications include opinion mining, human-computer interaction, plagiarism detection, and information retrieval, among many other tasks.

José Luis Oropeza Rodríguez, Sergio Suárez Guerra from México in their paper "Cochlear Mechanical Models used in Automatic Speech Recognition Tasks" write: In this paper we show that its possible unify two theories that we can find in the state of the art related with human hearing, one of them related with human perceptual phenomenon and the another one related with cochlear mechanic's models linear.

Tayfun Pay, Stephen Lucci, James L. Cox from United States in their paper "An Ensemble of Automatic Keyword Extractors: TextRank, RAKE and TAKE" write: We construct an ensemble method for automatic keyword extraction from single documents. We utilize three different unsupervised automatic keyword extractors in building our ensemble method. These three approaches provide candidate keywords for the ensemble method without using their respective threshold functions.

Armel Fotsoh, Annig Le Parc Lacayrelle, Christian Sallaberry from France in their paper "Complex Named Entities Extraction on the Web: Application to Social Events" write: In this paper, we focus on the extraction of social events in text from the web. We consider social events as complex Named Entities (NEs) i.e. NEs represented by a list of properties that can be simple values (text, number, etc.), "elementary" NEs and/or other complex NEs. Regarding the extraction of these complex NEs, our contribution focuses on the noisy context issue.

Abhilasha Sancheti, Natwar Modani, Gautam Choudhary, Chintha Priyadarshini, Sai Sandeep Moparthi from India in their paper "Understanding Blogs Through the Lens of Readers' Comments" write: In order to keep their audience engaged, authors need to make sure that the blogs or articles they write cater to the taste of their audience and are understood by them. With the rapid proliferation of online blogging websites, the participation of readers by expressing their opinions and reviews has also increased in the form of comments on the blogs.

Gokul S. Krishnan, Sowmya Kamath S. from India in their paper "Ontology-driven Text Feature Modeling for Disease Prediction using Unstructured Radiological Notes" write: Clinical Decision Support Systems (CDSSs) support medical personnel by offering aid in decision making and timely interventions in patient care. Typically such systems are built on structured Electronic Health Records (EHRs), which, unfortunately have a very low adoption rate in developing countries at present.

Altanbek Zulkhazhav, Zhanibek Kozhirbayev, Zhandos Yessenbayevy, Altynbek Sharipbay in their paper "Kazakh Text Summarization using

Fuzzy Logic" write: In this paper we present an extractive summarization method for the Kazakh language based on fuzzy logic. We aimed to extract and concatenate important sentences from the primary text to obtain its shorter form. With the rapid growth of information on the Internet there is a demand on its efficient and cost-effiective summarization.

Mohamed Ali Batita, Rami Ayadi, Mounir Zrigui from Tunisia their paper "Reasoning over Arabic WordNet Relations with Neural Tensor Network" write: Arabic WordNet is an important resource for many tasks of natural language processing. However, it suffers from many problems. In this paper, we address the problem of the unseen relationships between words in Arabic WordNet. More precisely, we focus on the ability for new relationships to be learned 'automatically' in Arabic WordNet from existing relationships.

Sanjay Kamath, Brigitte Grau, Yue Ma from France in their paper "Predicting and Integrating Expected Answer Types into a Simple Recurrent Neural Network Model for Answer Sentence Selection" write: Since end-to-end deep learning models have started to replace traditional pipeline architectures of question answering systems, features such as expected answer types which are based on the question semantics are seldom used explicitly in the models.

Marek Menšík, Marie Duží, Adam Albert, Vojtěch Patschka, Miroslav Pajr from Czech Republic in their paper "Refining Concepts by Machine Learning" write: In this paper we deal with machine learning methods and algorithms applied in learning simple con-cepts by their refining or explication. The method of re-fining a simple concept of an object O consists in discov-ering a molecular concept that defines the same or a very similar object to the object O. Typically, such a mo-lecular concept is a professional definition of the object, for instance a biological definition according to taxon-omy, or legal definition of roles, acts, etc.

Yingju Xia, Zhongguang Zheng, Yao Meng, Jun Sun from China in their paper "Semi-automatic Knowledge Graph Construction by Relation Pattern Extraction" write: Knowledge graphs represent information in the form of entities and relation-ships between them. A knowledge graph consists of multi-relational data, having entities as nodes and relations as edges.

Amarnath Pathak, Ranjita Das, Partha Pakray, Alexander Gelbukh from India in their paper "Extracting Context of Math Formulae contained inside scientific documents" write: A math formula present inside a scientific document is often preceded by its textual description, which is commonly referred to as the context of formula. Annotating context to the formula enriches its semantics, and consequently impacts the retrieval of mathematical contents from scientific documents.

Arda Akdemir, Tunga Gungor from Japan and Turkey in her paper "Joint Learning of Named Entity Recognition and Dependency Parsing using Separate Datasets" write: Joint learning of different NLP-related tasks is an emerging research field in Machine Learning. Yet, most of the recent models proposed on joint learning require a dataset that is annotated jointly for all the tasks involved.

Sunita Warjri, Partha Pakray, Saralin Lyngdoh, Arnab Kumar Maji from in India their paper "Identification of POS Tag for Khasi Language based on Hidden Markov Model POS Tagger" write: Computational Linguistic (CL) becomes an essential and important amenity in the present scenarios, as many different technologies are involved in making machines to understand human languages.

Christopher G. Harris, Padmini Srinivasan from United States in their paper "My Word! Machine versus Human Computation Methods for Identifying and Resolving Acronyms" write: Acronyms are commonly used in human language as alternative forms of concepts to increase recognition, to reduce duplicate

references to the same concept, and to stress important concepts.

Alejandra Segura Navarrete, Christian Vidal Castro, Clemente Rubio Manzano, Claudia Martínez Araneda from Chile in his paper "The role of WordNet similarity in the affective analysis pipeline" write: Sentiment Analysis (SA) is a useful and important discipline in Computer Science, as it allows having a knowledge base about the opinions of people regarding a topic. This knowledge is used to improve decision-making processes. One approach to achieve this is based on the use of lexical knowledge structures.

Rim Koulali, Abdelouafi Meziane from Morocco in their paper "A Comparative Study on Text Representation Models For Arabic Topic Detection" write: Topic Detection (TD) plays a major role in Natural Language Processing (NLP). Its applications range from Question Answering to Speech Recognition. In order to correctly detect document's topic, we shall first proceed with a text representation phase to transform the electronic documents contents into an efficiently software handled form.

Samarth Navali, Jyothirmayi Kolachalam, Vanraj Vala from India in their paper "Sentence Generation Using Selective Text Prediction" write: Text generation based on comprehensive datasets has been a well-known problem from several years. The biggest challenge is in creating a readable and coherent personalized text for specific user. Deep learning models have had huge success in the different text generation tasks such as script creation, translation, caption generation etc.

K. Chakma, Amitava Das, Swapan Debbarma from India in their paper "Deep Semantic Role Labeling for Tweets using 5W1H: Who, What, When, Where, Why and How" write: . In natural language understanding, Semantic Role Labeling (SRL) is considered as one of the important tasks and widely studied by the research community. State-of-the-art lexical resources have been in existence for defining the semantic role arguments with respect to the predicates.

Rejwanul Haque, Mohammed Hasanuzzaman, Arvind Ramadurai, Andy Way from Ireland in their paper "Mining Purchase Intent in Twitter" write: Most social media platforms allow users to freely express their beliefs, opinions, thoughts, and intents. Twitter is one of the most popular social media platforms where users' post their intent to purchase. A purchase intent can be defined as measurement of the probability that a consumer will purchase a product or service in future. Identification of purchase intent in Twitter sphere is of utmost interest as it is one of the most long-standing and widely used measures in marketing research.

Masaki Murata, Natsumi Morimoto from Japan in their paper "Relation between Titles and Keywords in Japanese Academic Papers using Quantitative Analysis and Machine Learning" write: In this study, we analyzed keywords from different academic papers using data from more than 300 papers. Using the concept of quantitative surveys and machine learning, we conducted various analyses on the keywords in different papers.

Sandra J. Gutiérrez Hinojosa, Hiram Calvo, Marco A. Moreno Armendáriz from Mexico in their paper "Central Embeddings for Extractive Summarization Based on Similarity" write: In this work we propose using word embeddings combined with unsupervised methods such as clustering for the multi-document summarization task of DUC (Document Understanding Conference) 2002.

Bharat Gaind, Nitish Varshney, Shubham Goel, Akash Mondal from India in their paper "Identifying Short-Term Interests from Mobile App Adoption Pattern" write: With the increase in an average user's dependence on their mobile devices, the reliance on collecting user's browsing history from mobile browsers has also increased. This browsing history is highly utilized in the advertising industry for providing targeted

ads in the purview of inferring user's short-term interests and pushing relevant ads.

Shubham Krishna, Ahsaas Bajaj, Mukund Rungta, Hemant Tiwari, Vanraj Vala from India in their paper "RelEmb: A Relevance-based Application Embedding for Mobile App Retrieval and Categorization" write: Information Retrieval Systems have revolutionized the organization and extraction of Information. In recent years, mobile applications (apps) have become primary tools of collecting and disseminating information.

Jakub Sido, Miloslav Konopík, Ondřej Pražák from Czech Republic in their paper "English Dataset for Automatic Forum Extraction" describe a novel dataset. They write: This paper describes the process of collecting, maintaining and exploiting an English dataset of web discussions. The dataset consists of many web discussions with hand-annotated posts in the context of a tree structure of a web page.

Imran Sheikh, Balamallikarjuna Garlapati, Srinivas Chalamala, Sunil Kumar Kopparapu from India in their paper "A Fuzzy Approach to Mute Sensitive Information in Noisy Audio Conversations" write: Audio conversation between a service seeking customer and an agent are common in a voice based call center (VbCC) and are often recorded either for audit purposes or to enable the VbCC to improve their efficiency. These audio recordings invariably contain personal information of the customer, often spoken by the customer to confirm their identity to get personalized services from the agent.

Girish K. Palshikar, Sachin Pawar, Rajiv Srivastava, Mahek Shah from India in their paper "Identifying Repeated Sections within Documents" write: Identifying sections containing a logically coherent text about a particular aspect is important for fine-grained IR, question-answering and information extraction. We propose a novel problem of identifying repeated sections, such as project details in resumes and different sports events in the transcript of a news broadcast. We focus on resumes and present four techniques (2 unsupervised, 2 supervised) for automatically identifying repeated project sections.

Alymzhan Toleu, Gulmira Tolegen, Rustam Mussabayev from Kazakhstan in their paper "KeyVector: Unsupervised Keyphrase Extraction Using Weighted Topic via Semantic Relatedness" write: . Keyphrase extraction is a task of automatically selecting topical phrases from a document. We present KeyVector, an unsupervised approach with weighted topics via semantic relatedness for keyphrase extraction.

Thuong-Hai Pham, Dominik Machácek, Ondřej Bojar from Czech Republic in their paper "Promoting the Knowledge of Source Syntax in Transformer NMT" write: The utility of linguistic annotation in neural machine translation has been already established. The experiments were however limited to recurrent sequence-to-sequence architectures and relatively small data settings.

Sapan Shah, Sarath S., Sreedhar Reddy from India in their paper "Similarity Driven Unsupervised Learning for Materials Science Terminology Extraction" write: Knowledge of material properties, microstructure, underlying material composition and manufacturing process parameters that the material has undergone is of significant interest to materials scientists and engineers.

Salima Harrat, Karima Meftouh, Kamel Smaili from Algeria and France in their paper "Script Independent Morphological Segmentation for Arabic Maghrebi Dialects: An Application to Machine Translation" write: This research deals with resources creation for under-resourced languages. We try to adapt existing resources for other resourced-languages to process less-resourced ones.

Lukas Svoboda, Tomas Brychcín from Czech Republic in their paper "Enriching Word Embeddings with Global Information and Testing

on Highly Inflected Language" write: In this paper we evaluate our new approach based on the Continuous Bag-of-Words and Skip-gram models enriched with global context information on highly inflected Czech language and compare it with English results. As a source of information we use Wikipedia, where articles are organized in a hierarchy of categories.

Rijul Dhir, Santosh Kumar Mishra, Sriparna Saha, Pushpak Bhattacharyya from India in their paper "A Deep Attention based Framework for Image Caption Generation in Hindi Language" write: Image captioning refers to the process of generating a textual description for an image which defines the object and activity within the image. It is an intersection of computer vision and natural language processing where computer vision is used to understand the content of an image and language modelling from natural language processing is used to convert an image into words in the right order.

Murillo Lagranha Flores, Elder Rizzon Santos, Ricardo Azambuja Silveira from Brazil in their paper "Ontology-based Extractive Text Summarization: The Contribution of Instances" write: In this paper, we present a text summarization approach focusing on multi-document, extractive and query-focused summarization that relies on an ontology-based semantic similarity measure, that specifically explores ontology instances.

Masahiro Kaneko, Mamoru Komachi from Japan in their paper "Multi-Head Multi-Layer Attention to Deep Language Representations for Grammatical Error Detection" consider the task of grammar correction using neural networks. They write: It is known that a deep neural network model pre-trained with large-scale data greatly improves the accuracy of various tasks, especially when there are resource constraints. However, the information needed to solve a given task can vary, and simply using the output of the final layer is not necessarily sufficient.

Basma El Amel Boussaha, Nicolas Hernandez, Christine Jacquin, Emmanuel Morin from France in their paper "Towards Simple but Efficient Next Utterance Ranking" consider the task of utterance ranking. They write: Retrieval-based dialogue systems converse with humans by ranking candidate responses according to their relevance to the history of the conversation (context). Recent studies either match the context with the response on only sequence level or use complex architectures to match them on the word and sequence levels. We show that both information levels are important and that a simple architecture can capture them effectively.

Shashavali D., Vishwjeet, Rahul Kumar, Gaurav Mathur, Nikhil Nihal, Siddhartha Mukherjee, Suresh Venkanagouda Patil from India in their paper "Sentence Similarity Techniques for short vs Variable Length Text using Word Embeddings" write: In goal-oriented conversational agents like Chatbots, finding the similarity between user input and representative text result is a big challenge. Generally, the conversational agent developers tend to provide a minimal number of utterances per intent, which makes the classification task difficult.

Lucie Poláková, Jiří Mírovský from Czech Republic in their paper "Anaphoric Connectives and Long-Distance Discourse Relations in Czech" write: This paper is a linguistic as well as technical survey for the development of a shallow discourse parser for Czech. It focuses on long-distance discourse relations signalled by (mostly) anaphoric discourse connectives.

Dwijen Rudrapal, Amitava Das, Baby Bhattacharya in their paper "A New Approach for Twitter Event Summarization Based on Sentence Identification and Partial Textual Entailment" write: Recent trend of information propagation on any real-time event in Twitter makes this platform more and more popular than any other online communication media. This trend creates a necessity to understand real-time events quickly and precisely by summarizing all the relevant tweets. In this paper, we propose a two-phase

summarization approach to produce abstract summary of any Twitter event.

Kushal Singla, Joy Bose, Nitish Varshney from India in their paper "Word Embeddings for IoT Based on Device Activity Footprints" write: With the expansion of IoT ecosystem, there is an explosion of the number of devices and sensors and the data generated by these devices. However, the tools available to analyze such data are limited. Word embeddings, widely used in the natural language processing (NLP) domain, provides a way to get similar words to the current word.

Nadeesha Pathirana, Sandaru Seneviratne, Rangika Samarawickrama, Shane Wolff, Charith Chitraranjan, Uthayasanker Thayasivam, Tharindu Ranasinghe from Sri Lanka in their paper "Concept Discovery through Information Extraction in Restaurant Domain" write: Concept identification is a crucial step in understanding and building a knowledge base for any particular domain.

Ajay Nagar, Anmol Bhasin, Gaurav Mathur from India in their paper "Text Classification using gated fusion of n-gram features and semantic features" consider a text classification task with feature fusion. They write: We introduce a novel method for text classification based on gated fusion of n-gram features and semantic features of the text. The parallel CNN network captures the n-gram relation between the words based on the filter size, primarily short distance multi-word relations.

Rodrigo López, Daniel Peñaloza, Francisco Beingolea, Juanjose Tenorio, Marco Sobrevilla Cabezudo from Brazil in their paper "An Exploratory Study of the Use of Senses, Syntax and Cross-Linguistic Information for Subjectivity Detection in Spanish" write: This work presents an exploratory study of Subjectivity Detection for Spanish This study aims to evaluate the use of dependency relations, word senses and cross-linguistic information in Subjectivity Detection task.

Lorenzo Porcaro, Horacio Saggion from Spain in their paper "Recognizing Musical Entities in User-generated Content" write: Recognizing Musical Entities is important for Music Information Retrieval (MIR) since it can improve the performance of several tasks such as music recommendation, genre classification or artist similarity.

Maria Janicka, Maria Pszona, Aleksander Wawer from Poland in their paper "Cross-Domain Failures of Fake News Detection" consider the task of fake news detection. In their paper, they write: Fake news recognition has become a prominent research topic in natural language processing. Researchers reported significant successes when applying methods based on various stylometric and lexical features and machine learning, with accuracy reaching 90%. This article is focused on answering the question: are the fake news detection models universally applicable or limited to the domain they have been trained on? We used four different, freely available English language Fake News corpora and trained models in both in-domain and cross-domain setting. We also explored and compared features important in each domain. We found that the performance in cross-domain setting degrades by 20% and sets of features important to detect fake texts differ between domains. Our conclusions support the hypothesis that high accuracy of machine learning models applied to fake news detection may be related to over-fitting, and models need to be trained and evaluated on mixed types of texts.

Mohammed Alliheedi, Robert E. Mercer, Sandor Haas Neil from Canada in their paper "Ontological Knowledge for Rhetorical Move Analysis" write: Scholarly writing in the experimental biomedical sciences follows the IMRaD (Introduction, Methods, Results, and Discussion) structure. Many Biomedical Natural Language Processing tasks take advantage of this structure. Recently, a new challenging information extraction task has been introduced as a means of obtaining these types of detailed

information: identifying the argumentation structure in biomedical articles. Argumentation mining can be used to validate scientific claims and experimental methodology, and to plot deeper chains of scientific reasoning. One subtask in identifying the argumentation structure is the identification of rhetorical moves, text segments that are rhetorical and perform specific communicative goals, in the Methods section. Based on a descriptive taxonomy of rhetorical moves structured around IMRaD, the foundational linguistic knowledge needed for a computationally feasible model of the rhetorical moves is described: semantic roles. One goal is to provide FrameNet and VerbNet-like ontologies for the specialized domain of biochemistry. Using the observation that the structure of scholarly writing in the laboratory-based experimental sciences closely follows the laboratory procedures, we focus on the procedural verbs in the Methods section. Occasionally, the text does not contain fillers for all of the semantic role slots that are needed to perform an adequate analysis of a verb. To overcome this problem, an ontology of experimental procedures can be interrogated to provide a most likely candidate for the missing semantic role(s).

The issue will be of interest for students and researchers working in natural language processing, computational linguistics and related areas of artificial intelligence and data science.

Alexander Gelbukh
Guest Editor

Instituto Politécnico Nacional,
Centro de Investigación en Computación,
Mexico City, Mexico