ISSN 2007-9737

Introduction to the Thematic Section on Computational Linguistics

This thematic section of *Computación y Sistemas* presents a selection of papers on computational linguistics.

Computational linguistics is a very actively developed research area at the crossroads of linguistics, computer science, and artificial intelligence, thus embracing theoretical research and its practical application and implementation issues. Its main applications are aimed to analysis and generation of the data that reflect human communication using language, such as English or Spanish, in the form of text or speech. Computational linguistics, which is the main theoretical basis behind natural language processing, enables the use of computers in situations related to human communication in a wider context of their real-life applications and in a wide range of disciplines.

Computational linguistics addresses many computational tasks related with natural language, such as machine translation, text classification, text summarization, information extraction, and sentiment analysis, while its typical applications include opinion mining, human-computer interaction, and information retrieval, among many other tasks.

A fascinating feature of computational linguistics research, as compared with other computer science tasks, is its application to very different human languages, with different morphological and syntactic types and different writing systems. For example, in this thematic section the reader will find papers that analyze applications of computational linguistics techniques to Russian, Vietnamese, Indian languages including even Sanskrit, and many others.

Another interesting feature of the modern computational linguistics research is active use of machine-learning techniques, which lay out solid mathematical foundation for the computational methods used in those works. The reader will find quite a number of interesting machine-learning techniques developed and improved in this thematic section.

For the convenience of the reader, in the sequel, the contents of the papers included in this thematic section is briefly summarized.

Carlos A. Rodriguez-Diaz, Sergio Jimenez, George Dueñas, Johnatan Estiven Bonilla, and Alexander Gelbukh from Colombia and Mexico in their paper “Dialectones: Finding Statistically Significant Dialectal Boundaries Using Twitter Data” adapt the concept of “ecotone” to “dialectone” for the detection of dialectal boundaries by using two non-parametric statistical tests: the Hilbert-Schmidt independence criterion (HSIC) and the Wilcoxon signed-rank. Their method was applied to a large corpus of Spanish tweets produced in 160 locations in Colombia through the analysis of unigram features. The resulting dialectones showed to be meaningful but difficult to compare against regions identified by other authors using classical dialectometry. They concluded that the automatic detection of dialectones is convenient alternative to classical methods in dialectometry and a potential source of information for automatic language applications.

Hala Mulki, Hatem Haddad, Chedi Bechikh Ali, and Ismail Babaoglu from Turkey, Belgium, and Tunisia in their paper “Tunisian Dialect Sentiment Analysis: A Natural Language Processing-based Approach” investigate the impact of several preprocessing techniques on sentiment analysis using two sentiment classification models: supervised and lexicon-based. These models were trained on three Tunisian datasets of different sizes and multiple domains. Their results emphasize the positive impact of preprocessing phase on the evaluation measures of both sentiment classifiers as the baseline was significantly outperformed when stemming, emoji

doi: 10.13053/CyS-22-4-3106

recognition and negation detection tasks were applied. Moreover, integrating named entities with these tasks enhanced the lexicon-based classification performance in all datasets and that of the supervised model in medium and small sized datasets.

Marta R. Costa-jussà, Álvaro Nuez, and Carlos Segura from Spain in their paper "Experimental Research on Encoder-Decoder Architectures with Attention for Chatbots" offer an experimental view of how recent advances in such areas as machine translation can be adopted for chatbots. In particular, they compare how alternative encoder-decoder deep learning architectures perform in the context of chatbots. Their research concludes that a fully attention-based architecture is able to outperform the recurrent neural network baseline system.

Ankush Khandelwal, Sahil Swami, Syed Sarfaraz Akhtar, and Manish Shrivastava from India in their paper "Gender Prediction in English-Hindi Code-Mixed Social Media Content: Corpus and Baseline System" analyze the task of author's gender prediction in code-mixed content and present a corpus of English-Hindi texts collected from Twitter which is annotated with author's gender. They also explore language identification of every word in this corpus. They present a supervised classification baseline system, which uses various machine-learning algorithms to identify the gender of an author using a text, based on character and word level features.

Gaël Lejeune and Lichao Zhu from France in their paper "A New Proposal for Evaluating Web Page Cleaning Tools" tackle the problem of evaluation of Web Content Extraction tools. This task is seldom studied in the literature although it has important consequences on the linguistic processing of web-based corpora. Here, they compare two types of evaluation. First, an intrinsic (content-based) evaluation which is carried out in a multilingual setting (five languages). Second, an extrinsic (task-based) evaluation on the same corpus by studying the effects of the cleaning step on the performances of an NLP pipeline. They show that in the intrinsic evaluation, the results are not consistent with extrinsic evaluation results. They also show that the results differ greatly in the studied languages. They conclude that the choice of a web page cleaning tool should be made with respect to the task that is tackled rather than the performances observed through the intrinsic evaluation scheme.

Renata Vieira, Amália Mendes, Paulo Quaresma, Evandro Fonseca, Sandra Collovini, and Sandra Antunes from Brazil and Portugal in their paper "Corref-PT:A Semi-Automatic Annotated Portuguese Coreference Corpus" describe the Portuguese coreference corpus Corref-PT, annotated semi-automatically using the coreference annotation tool CORP, and manually revised with the editing tool CorrefVisual. It includes a total of 182 texts, mostly news (corpus CSTNews, corpus LE-PAROLE, FAPESP magazine) but also articles from Wikipedia. The result is a corpus that includes a total of 3898 reference chains. They present the coreference annotation tool CORP, which was built on the basis of deterministic rules, and the editor CorrefVisual used for manual revision. They report on the annotation agreement and on the feedback provided by the annotators regarding the editor and the complexity of the task. Examples of technical and linguistic issues encountered during the annotation are given and the pros and cons of such approach for corpus construction are discussed. Their motivation was to use of a semi-automatic approach to increase the set of available resources for coreference resolution applications for Portuguese.

Hiram Calvo from Mexico in his paper "Psychological Attachment Style Determination Using a Word Space Model" experiments with the hypothesis that words a subject uses and their psychological attachment style (as defined by Bartholomew and Horowitz), can be related. In order to verify this hypothesis, he identified characteristic patterns for each style of attachment (secure, fearful, dismissing, preoccupied) by mapping words into a word space model on a series of autobiographic texts written by a set of 202 participants. Additionally, a psychological instrument (questionnaire) was applied to these same participants to measure their attachment style. A Support Vector Machine was trained, and he found that attachment style could be predicted from text within a range of 64% to 85% for different attachment styles.

Vijay Kumar Sharma and Namita Mittal from India in their paper "An Improvement in Statistical Machine Translation in Perspective of Hindi-English Cross-Lingual Information Retrieval" train a Statistical Machine Translation (SMT) system on two parallel corpora separately. A large English language corpus is used for language modeling in SMT. Experiments are evaluated by using BLEU score, further, these experimental setups are used to translate the Hindi language queries for the experimental analysis of Hindi-English CLIR. Since SMT does not deal with morphological variants while the proposed Translation Induction Algorithm (TIA) deals with that, therefore, TIA outperforms the SMT systems in perspective of CLIR.

Hiep Nguyen Minh, Huyen Nguyen Thi Minh, and Quyen Ngo The from Vietnam in their paper "Building Resources For Vietnamese Clinical Text Processing" investigate the tasks of lexical analysis and phrase chunking for Vietnamese clinical texts. Although there exist several tools for general Vietnamese text analysis, these tools showed a limited quality in the clinical domain due to the specific grammatical style of clinical texts and the lack of medical vocabulary. Their main contributions are the construction of an annotated corpus (vnEMR) and lexical resources in the medical domain and in consequence the improvement of the quality of the tools for clinical text analysis, including word segmentation, part-of-speech tagging and chunking.

Jon Alkorta, Koldo Gojenola, and Mikel Iruskieta from Spain in their paper "SentiTegi: Semi-manually Created Semantic Oriented Basque Lexicon for Sentiment Analysis" present the construction and evaluation of the first semantic oriented supervised Basque lexicon ranging from +5 to −5. Due to the lack of resources, the Basque lexicon was created translating the SO-CAL Spanish dictionary by means of two bilingual dictionaries following specific criteria and then slightly corrected with the SO-CAL English dictionary and frequency data obtained from the Basque Opinion Corpus. Evaluation results show that the correlation between human annotators is slightly better than between a gold standard lexicon (obtained from human annotation) and the translated dictionary. This shows that the quality of the translated lexicon is satisfactory, although there is a space to improve it.

Ayush Aggarwal, Chhavi Sharma, Minni Jain, and Amita Jain from India in their paper "Semi Supervised Graph Based Keyword Extraction Using Lexical Chains and Centrality Measures" present keyword extraction using lexical chains and graph centrality measures, derived from the semantic similarity of the words by analysis of the graphical network created using WordNet. The hypothesis is presented using a small-world approach where every paragraph in a document is constrained to a local point, while the document in all is centered on a global concept. Creating lexical chains for each paragraph and combining the best via scoring methods and graph-based algorithms, we present parallels to baseline system to extract the keywords from the document.

Tu Vu, Xuan Bui, Khoat Than, and Ryutaro Ichise from Vietnam and Japan in their paper "A Flexible Stochastic Method for Solving the MAP Problem in Topic Models" propose a more general and flexible version of OPE, namely Generalized Online Maximum a Posteriori Estimation (G-OPE), which not only enhances the flexibility of OPE in different real-world situations but also preserves key advantage theoretical characteristics of OPE when comparing to the state-of-the-art methods. They employ G-OPE as inference a document within large text corpora.

The experimental and theoretical results show that their new approach performs better than OPE and other state-of-the-art methods.

Elena Yagunova, Ekaterina Pronoza, and Nataliya Kochetkova from Russia in their paper “Construction of Paraphrase Graphs as a Means of News Clusters Extraction” construct paraphrase graphs for news text collections (clusters). Their aims are, first, to prove that paraphrase graph construction method can be used for news clusters identification and, second, to analyze and compare stylistically different news collections. Their news collections include dynamic, static and combined (dynamic and static) texts. The respective paraphrase graphs reflect their main characteristics. They also automatically extract the most informationally important linked fragments of news texts, and these fragments characterize news texts as either informative, conveying some information, or publicistic ones, trying to affect the readers emotionally.

Amal Rekik, Hanen Ameur, Amal Abid, Atika Mbarek, Wafa Kardamine, Salma Jamoussi, and Abdelmajid Ben Hamadou from Tunisia in their paper “Building an Arabic Social Corpus for Dangerous Profile Extraction on Social Networks” propose a new method for data extraction and annotation of suspicious users from social networks threatening the national security. Their method allows constructing a rich Arabic corpus designed for detecting terrorist users spreading on social networks. The amendment of our corpora is ensured following a set of rules defined by a domain expert. All these steps are described in details, and some typical examples are given. Also, some statistics are reported from the data collection and annotation stages as well as the evaluation of the annotated features based on the intra-agreement measurement between different experts.

Dror Mughaz, Tzeviya Fuchs, and Dan Bouhnik from Israel in their paper “Automatic Opinion Extraction from Short Hebrew Texts Using Machine Learning Techniques” focus on classifying Modern Hebrew sentences according to their polarity. They compare various Machine Learning algorithms and techniques of classification. They added optimizations and methods that have not previously been used, and adjusted commonly used techniques so they would suit a Hebrew corpus. The authors elaborate on the differences in classifying short texts versus long ones and about the uniqueness of working specifically with Hebrew. Finally, their model achieved nearly 93% accuracy, which is higher than accuracies achieved previously in this field.

Lokesh Kumar Sharma, Namita Mittal, and Anubha Aggarwal from India in their paper “Feature Extraction for Token Based Word Alignment for Question Answering Systems” develop an aligner which despite using very little lexical resources gives good results in terms of precision, recall, and F1. Previous aligners either uses more lexical resources or uses very less lexical resources. Hence, they have used POS TAG and WordNet as lexical resources. However, some words whose meaning we may not know but these occur in a similar distribution and by observing their distribution these words are similar. For example, in the two sentences “Lambodar is the son of Parvati” and “Ganesha is the son of Parvati”, one will not find the meaning of Lambodar and Ganesha in Wordnet but since they have similar distributions so they should be aligned. For these words, the authors used Distribution Similarity Feature in their word aligner. This distributional similarity helps their aligner in broader coverage of words. Previous aligners were having recall in the range of 75 to 86, but this aligner has recall in the range of 88.4 to 93.3. Similarly, exact match of previous aligners was in the range of 21-35.3 but the proposed aligner’s exact match range is 46.1 to 58.6. Similarly, F-measure and precision have increased.

Muhammad Rif’at, Rahmad Mahendra, Indra Budi, and Haryo Akbarianto Wibowo from

Indonesia in their paper “Towards Product Attributes Extraction in Indonesian e-Commerce Platform” present a study of attribute extraction from Indonesian e-commerce product titles. They annotate 1,721 product titles with 16 attribute labels. They apply supervised learning technique using CRF algorithm. They propose combination of lexical, word embedding, and dictionary features to learn the attribute using joint extraction model. Their model achieves F1-measure 47.30% and 68.49% respectively for full match and partial match evaluation. Based on the experiment, they find that doing attributes extraction on more various number and diverse attributes simultaneously does not necessarily give worse result compared to extraction on less number of attributes.

Puneet Dwivedi and Daniel Zeman from Czechia in their paper “The Forest Lion and the Bull: Morphosyntactic Annotation of the Panchatantra” present the first freely available dependency treebank of Sanskrit. It is based on text from Panchatantra, an ancient Indian collection of fables. The annotation scheme we chose is that of Universal Dependencies, a current de-facto standard for cross-linguistically comparable morphological and syntactic annotation. In this paper, they discuss word segmentation issues, morphological inventory and certain interesting syntactic constructions in the light of the Universal Dependencies guidelines. They also present an initial parsing experiment.

Thi-Lan Ngo Thi-Lan Ngo, Tu Vu, Hideaki Takeda, Son Bao Pham, and Xuan Hieu Phan from Vietnam and Japan in their paper “Lifelong Learning Maxent for Suggestion Classification” introduce a method called LLMaxent which is the solution for the cross-domain suggestion classification. LLMaxent is a lifelong machine learning approach using maximum entropy (Maxent). In the course of lifelong learning, the drawn knowledge from the past tasks is retained and supported for the future learning. From that, we build a classifier by using labeled data in existed domains for suggestion classification in a new domain. The experimental results show that the proposed novel model can improve the performance of cross-domain suggestion classification. This is one of the preliminary research in lifelong machine learning using Maxent. Its effect is not only for suggestion classification but also for cross-domain text classification in general.

Hirotaka Niitsuma from Japan in his paper “Context-Free Grammars Including Left Recursion using Recursive miniKanren” define a pattern match macro which can use the same syntax of the match macro of the Scheme language using recursive miniKanren. The macro enables to write searching sub-list with a given pattern by only few line code. Using this property, he introduces techniques writing context-free grammar with his match macro. Unlike other specific paraphrasing tools, his technique can combine logical relations of miniKanren with a context-free grammar. He shows the logical relations resolves the ambiguity of a grammar.

Kavita Asnani and Jyoti D. Pawar from India in their paper “Improving Coherence of Topic Based Aspect Clusters using Domain Knowledge” use context domain knowledge from a publicly available lexical resource to increase the coherence of topic-based aspect clusters and discriminate domain-specific semantically relevant topical aspects from generic aspects shared across the domains. BabelNet was used as the lexical resource. The dataset comprised of product reviews from 36 product domains, containing 1000 reviews from each domain and 14 clusters per domain. Also, frequent topical aspects across topic clusters indicate occurrence of generic aspects. The average elimination of incoherent aspects was found to be 28.84%. The trend generated by UMass metric shows improved topic coherence and also better cluster quality is obtained as the average entropy without eliminated values was 0.876 and with elimination was 0.906.

Sana Fakhfakh and Yousra Ben Jemaa from Tunisia in their paper "Gesture Recognition System for Isolated Word Sign Language Based on key-Point Trajectory Matrix" suggest a new system to help the deaf and the hearing-impaired community improve their connection with the hearing world and communicate freely. The most important thing in this system is how to help the users be free and finally have a more natural way of communication. For this reason, they present a new process based on two levels: a static-level aiming to extract the most head/hands key points and a dynamic-level with the objective of accumulating the key-point trajectory matrix. Also their proposed approach takes into account the signer-independence constraint. A SIGNUM database is applied in the classification stage and their system performance has improved with a 94.3% recognition rate. Furthermore, a reduction in time processing is obtained when the removing of redundant frame step is applied. The obtained results prove the superiority of their system compared to the state-of-the-art methods in terms of recognition rate and execution time.

The papers included in this thematic section will be useful to researchers and students who use natural language processing and computational linguistics techniques in their work. Some of the papers in this thematic section also give good samples of the use of machine-learning methods in sequential data analysis and specifically in natural language processing.

Alexander Gelbukh (Guest editor)

Member of the Mexican Academy of Sciences;
Head, Natural Language Processing Laboratory,
Centro de Investigación en Computación,
Instituto Politécnico Nacional, Mexico