NLPCR18

II Coloquio Costarricense de Procesamiento del Lenguaje Natural

Actas del Coloquio

Universidad de Costa Rica

25 y 26 de octubre del 2018.

Montes de Oca, San Pedro, Costa Rica.

Organizadores del Evento:

Daniela Sánchez Sánchez

Edgar Casasola Murillo

Éricka Vargas Castro

Federico Pacheco Rivera

Gabriela Marín Raventós

Jorge Antonio Leoni de León

Rocío Ruiz Ramón

Comité Científico

Aurelio Sanabria Rodríguez (ITCR)

Alfonso Ureña López (Sociedad Española para el procesamiento del Lenguaje Natural (SEPLN))

César Aguilar

Edgar Casasola Murillo (CITIC / ECCI)

Eugenio Martínez Cámara

Gabriela Marín Raventós (CITIC / ECCI)

Gabriela Barrantes Sliesarieva (CITIC / ECCI)

Gerardo Sierra Martínez (Grupo de Ingeniería Lingüística (GIL) , Universidad Autónoma de México)

Jorge Antonio Leoni de León (INIL)

Mario Hernández Delgado (INIL/ELEXHICÓS)

Sergio Cordero Monge (INIL/ELEXHICÓS)

Artículos aceptados

Nota

Debido a que los autores han enviado sus propuestas a revistas indexadas, únicamente publicamos en estas actas un resumen de las ponencias

Comparación de tres modelos de redes neuronales profundas para clasificar la polaridad de Tweets en español

Comparison of three deep learning neural networks for classifying polarity of spanish Tweets

Esteban Rodríguez Betancourt, Pablo Sauma Chacón, Edgar Casasola Murillo
Sistema de Estudios de Posgrado, Universidad de Costa Rica
estebarb@gmail.com, pablo.saumachacon@ucr.ac.cr, edgar.casasola@ucr.ac.cr

Resumen: Con el propósito de clasificar la polaridad o sentimiento de publicaciones en español en redes sociales se evaluaron tres modelos diferentes de redes neuronales profundas con arquitecturas diferentes. Para los tres modelos se calculó la exactitud, precisión, *recall* y F1 al ser entrenados sobre el corpus general de TASS. Se obtuvieron resultados exitosos en la clasificación, aunque se detectaron características de los textos que aún no son bien interpretadas por estas técnicas de clasificación.
Palabras clave: análisis de sentimiento, redes neuronales, deep learning

Abstract: With the purpose of classifying polarity in Spanish tweets three different deep learning models were evaluated. Accuracy, precision, recall and F1 were calculated for the three models, trained over general TASS Corpus. Although good results were achieved, some text aspects were identified that still cause some trouble to this classification techniques.
Keywords: sentiment analysis, neuronal networks, deep learning

Aplicación de Deep learning para análisis de sentimiento: Un análisis de comentarios en español de Twitter

Deep learning application for sentiment analysis: An analysis of spanish comments in Twitter

José-Joaquín Peralta-Abadía[1], **Luis Sánchez-Vargas**[2], **Carlos Solís-Fonseca**[3], **Edgar Casasola-Murillo**[2]
[1]josejoaquin.peralta.abadia@gmail.com
[2]lm.sanchezvargas@gmail.com
[3]solfonca@gmail.com
[4]edgar.casasola@ucr.ac.cr
Universidad de Costa Rica

Resumen: En este artículo se presenta un clasificador basado en una redes neuronales convolucional (CNN) aplicado a la Tarea 1 del taller TASS 2017 de clasificación de sentimientos de tweets. Utilizamos técnicas de pre-procesamiento y generamos vectores de palabras a partir de los los tweets para alimentárselos al clasificador. Se presentan los resultados preliminares obtenidos de la evaluación de un segmento del corpus de entrenamiento. Como trabajo futuro, pensamos incorporar *lexicons* de sentimientos y usar vectores de palabras específicos de sentimientos.
Palabras clave: Vectores de palabras, Análisis de sentimientos, Procesamiento del lenguaje natural, Aprendizaje profundo, Twitter

Abstract: In this article we present a classifier based on a convolutional neuronal network (CNN) applied to Task 1 of the TASS 2017 workshop for sentiment classification of tweets. We used pre-processing techniques and generated word embeddings from the tweets and fed them to the classifier. The preliminary results obtained from the evaluation of a segment of the training corpus are presented. As future work, we think of incorporating sentiment lexicons and using sentiment specific work embeddings.
Keywords: Word embeddings, Sentiment analysis, Natural language processing, Deep learning, Twitter

Uso de vectores de palabras para análisis de sentimientos en español: una revisión de literatura

Use of word embeddings for sentiment analysis in Spanish: a literature review

José-Joaquín Peralta-Abadía[1], **Edgar Casasola-Murillo**[2]
[1]josejoaquin.peralta.abadia@gmail.com
[2]edgar.casasola@ucr.ac.cr
Universidad de Costa Rica

Resumen: Este estudio exploratorio pretende identificar las técnicas de uso y limitantes de los *word embeddings* o vectores de palabras en tareas de análisis de sentimientos con énfasis en español. Para realizar este estudio se usaron como fuentes a *IEEE Xplore*, *SpringerLink*, *ACM Digital Library* y *Google Scholar*. Como criterios de elegibilidad se valida que tengan fechas del 2014 en adelante, que sean en textos en idioma inglés y español, y que provengan de conferencias, libros, revistas o actas. De los resultados se pudo identificar que el estado del arte entre español e inglés tiene similar nivel de avance, en cuanto a las técnicas utilizadas para generar y usar los vectores de palabras. El uso de vectores de palabras en español tiene varias configuraciones que impactan los resultados de forma positiva y negativa, por lo que más investigación es requerida.
Palabras clave: Vectores de palabras, Análisis de sentimientos, Procesamiento del lenguaje natural

Abstract: This exploratory study aims to identify the techniques of use and limitations of word embeddings in tasks of sentiment analysis with emphasis on Spanish. To carry out this study, we used IEEE Xplore, SpringerLink, ACM Digital Library and Google Scholar as sources. It is validated that they have dates from 2014 onwards, which are in texts in English and Spanish, and that come from conferences, books, magazines or proceedings as eligibility criteria. From the results it was possible to identify that the state of the art between Spanish and English has a similar level of progress, regarding the techniques used to generate and use word embeddings. The use of word embeddings in Spanish has several configurations that impact the results in a positive and negative way, so more research is required.
Keywords: Word embeddings, Sentiment analysis, Natural language processing

Consideraciones sobre el uso de modelos distribucionales en el contexto del estudio lingüístico

Some considerations regarding the use of word embeddings in the context of linguistics research

Federico Pacheco[1]
[1]Universidad de Costa Rica

Resumen: En este artículo se evalúan tres aspectos del uso de representaciones vectoriales de palabras que deben ser tomados en cuenta si se pretende utilizar estos modelos distribuidos en investigaciones lingüísticas: la forma de evaluar el modelo, la parametrización del entrenamiento, y el procesamiento de los datos base.
Palabras clave: Semántica, modelos semánticos distribucionales, metodología lingüística

Abstract: In this article, we evaluate three key aspects that must be taken into account if word vectors are to be used in linguistics research: model evaluation methodology, hyper-parameters of the training algorithms, and base data pre-processing.
Keywords: semantics, distributional semantic models, linguistics methodology

Desarrollo de un analizador morfológico de la lengua bribri con base en el modelo de estados finitos

Development of a morphological analyzer of the Bribri language based on the finite state model

Sofía Flores-Solórzano[1]

[1] https://orcid.org/0000-0003-3607-8662

Resumen: El bribri es una lengua de la estirpe chibchense que hablan aproximadamente unos 7.000 indígenas en el sudeste de Costa Rica. Nuestra meta ha sido desarrollar un analizador morfológico automático para esta lengua a partir de las técnicas de los estados finitos (Beesley y Karttunen, 2003; Jurafsky y Martin, 2008). Para ello hemos definido la morfotaxis del bribri, es decir, el orden de los morfemas y las combinaciones permitidas, así como una serie de reglas que contienen las alteraciones fonológicas y ortográficas necesarias. Tras compilar conjuntamente el lexicón y las reglas de alteración obtenemos un transductor de estados finitos que funciona como analizador y generador morfológico de esta lengua. La herramienta se encuentra disponible en la URL http://morphology.bribri.net/.
Palabras clave: bribri, lenguas indígenas, morfología, máquina de estados finitos, análisis morfológico.

Abstract: Bribri is a Chibchan language spoken by 7,000 people in Costa Rica. Our foremost aim has been to apply finite states techniques (Beesley y Karttunen, 2003; Jurafsky y Martin, 2008) to develop an automatic morphological analyzer for Bribri language. To do so we defined the bribri morphotaxis, i.e., the order of the morphemes and their legal combinations and a series of rules with the phonological and orthographic alterations. After the compilation of the lexicon and the rules of alteration we obtain a finite-state transducer that functions as a morphological analyzer and generator of this language. This tool is available at the URL http://morphology.bribri.net/
Keywords: Bribri, Indigenous languages, morphology, finite state machine, morphological analyzer.

Implementación de Latent Dirichlet Allocation LDA en Apache Spark para análisis de tópicos un caso de estudio sobre interacciones en chats de videojuegos para todas las edades

Implementation of Latent Dirichlet Allocation LDA with Apache Spark for topic analysis a case study on chats interactions in video games rated for all ages

Noelia Navarro[1], Paulo Calvo[2]
Universidad de Costa Rica
{noelia.navarromurillo[1], paulo.calvo[2]}@ucr.ac.cr

Resumen: Este artículo se enfoca en los resultados obtenidos en la implementación de la librería de Machine Learning de Apache Spark para el análisis de hilos temáticos de conversaciones extraídas de chats sobre videojuegos. Las comunidades en línea destinadas a discutir acerca de videojuegos representan una exposición para los niños y adolescentes respecto a temas y lenguaje utilizado en estas. A fin de conocer las temáticas abordadas en los chats procedimos a preprocesar el texto, agrupar interacciones e implementar el algoritmo generativo de agrupación Latent Dirichlet Allocation LDA. Este estudio provee información sobre los desafíos en el procesamiento de mensajes procedentes de interacciones de chats y los resultados obtenidos en la implementación del LDA.
Palabras clave: Latent Dirichlet Allocation, Apache Spark, video juegos, procesamiento de lenguaje natural.

Abstract: This article focuses on the results obtained in the implementation of the Apache Spark machine learning library for the analysis of thematic threads of conversations extracted from video game chats. The online communities destined to discuss about videogames represent an exhibition for children and adolescents regarding themes and language used in them. In order to know the topics addressed in the chats, we proceeded to preprocess the text, group interactions and implement the generative algorithm of the Latent Dirichlet Allocation LDA grouping. This study provides information on the challenges in the processing of messages from chat interactions and the results obtained in the implementation of the LDA
Keywords: Latent Dirichlet Allocation, Apache Spark, videogames, natural language processing

Procesamiento de textos de contenido comunitario centroamericano

Processing of Central American Spanish community content texts

Randall Araya Campos[1], **Paula Estrella**[2]
[1]Escuela de Ingeniería en Computación, Instituto Tecnológico de Costa Rica
[2]Facultad de Lenguas, Universidad Nacional de Córdoba
raraya@ic-itcr.ac.cr[1], pestrella@gmail.com[2]

Resumen: En este trabajo nos concentramos en estudiar la creación de una lista de palabras frecuentes del español centroamericano. Un aspecto novedoso de este trabajo es la utilización de corpus relacionados a la educación básica centroamericana *Escuela para todos*, y los aportes más significativos son: la creación de una lista de frecuencia de palabras abierta (opensource) y la propuesta de utilizar esta lista en conjunto con métricas para medir la complejidad textual de forma automatizada.
Palabras clave: español centroamericano, procesamiento de contenido educativo, escuela para todos, lenguaje comunitario

Abstract: In this paper we concentrate on studying the creation of a list of the vocabulary in Spanish of the Central America. A novel aspect of this work is the use of corpus related to basic education in Central America School for all, and the most significant contributions are: the creation of an open list of words, and the proposal to use this list in conjunction with metrics to measure textual complexity in an automated way.
Keywords: central american spanish, educational content processing, school for all, community language

Marcadores epistémicos en el género trabajo final de grado en español

Epistemic markers in an undergraduate graduation project in Spanish

Enrique Sologuren[1], René Venegas[2]
[1] Pontificia Universidad Católica de Valparaíso, Chile
[2] Pontificia Universidad Católica de Valparaíso, Chile

(enrique.sologuren.i@mail.pucv.cl ; rene.venegas@pucv.cl)

Resumen: En el presente trabajo se describen las estructuras lingüísticas de nivel léxico que expresan la modalidad epistémica en un corpus de Trabajos Finales de Grado (TFG) en español. La modalidad epistémica corresponde a categoría semántica-discursiva que expresa grados de certeza en los textos. Los datos muestran una mayor predominancia de los marcadores de modalidad epistémica no asertiva frente a los marcadores de modalidad epistémica asertiva. Los marcadores de modalidad epistémica *asertiva* predominan en el cumplimiento de los propósitos comunicativos "presentar investigaciones previas y antecedentes conceptuales relevantes (MM2)" y "Dar cuenta de los resultados y su interpretación en el contexto de la investigación (MM4)" al igual que los marcadores *no asertivos*, pero con distinciones disciplinares relevantes. Los hallazgos permiten comprender la forma en que cada disciplina construye de forma diversa el posicionamiento en el discurso académico estudiantil y ser potenciales atributos de clasificación de propósitos discursivos del género TFG.

Palabras clave: Modalidad epistémica, marcadores del discurso, trabajo final de grado, discurso académico.

Abstract: Linguistic structures of lexical level that show the epistemic modality in a corpus of Undergraduate Graduation Projects (TFG in Spanish) written in Spanish are described in this paper. The epistemic modality belongs to the discursive-semantic category that expresses degrees of certainty in those texts. Data show more predominance of non-assertive epistemic modality markers compared to assertive epistemic modality markers. These last markers predominate in accomplish the communicative purposes "to show previous research and important conceptual background (MM2)" and "to show results and their interpretation in the research context(MM4)" as well as non-assertive markers, however with important disciplinar distinctions. Findings help to understand the way each discipline constructs its positioning differently in the student academic discourse, and to be possible attributes of classification of discursive purposes in the TFG genre.

Keywords: Epistemic modality, discourse markers, undergraduate graduation project, academic discourse.