20 AÑOS Ingeniería Lingüística

IX CoLiCo

Coloquio de Lingüística Computacional

Septiembre
3 y 4
2019

Programa de Actividades

Salón de Actos, Facultad de Filosofía y Letras de la UNAM

www.colico.unam.mx

INSTITUTO DE INGENIERÍA UNAM
Facultad de Medicina
FACULTAD DE FILOSOFÍA Y LETRAS

PROGRAMA

Martes 3 de septiembre de 2019

Mesas	Horario	Ponencias	Autores
Inauguración	9:45 - 10:10		
Mesa 1 Lingüística Forense **Moderadora: Ana Aguilar Guevara**	10:20-10:40	Valores acústicos de referencia para el análisis forense de	Yunuen Sarasuadi Acosta Meza, Iván Vladimir Meza Ruíz, Fernanda López Escobedo / ENAH, IIMAS, LCF
	10:40 -11:00	Atribución de autoría mediante aprendizaje supervisado de características estilométricas	Tonatiuh Hernández García / GIL - IINGEN
	11:00- 11:20	#NiUnaMenos: violencia de género en los microdiscursos de Twitter	Michelle Denise Rodríguez Chiw, Tonatiuh Hernández / García Posgrado en Ciencia e Ingeniería de la Computación - UNAM, Licenciatura en lenguas y literaturas Hispánicas - UNAM.
	11:20 - 11:40	Programa BioMetro Fore como una alternativa para el análisis forense	Carlos Misael González Valseca, Sergio Arturo Dominguez Ortega / SUEDIF-CIESAS y ENAH.
	11:40 - 11:55	Preguntas y respuestas	
Ponencia Magistral		Ingeniería Lingüística en la UNAM a 20 años	Gerardo Sierra Martínez / GIL-IINGEN
Mesa 2. Enseñanza **Moderadora: Helena Gómez Adorno**	12:50 -13:10	Asistente computacional para el diseño de instrumentos de evaluación de la comprensión de lectura	Carlos César Jiménez, José Luis Sánchez García / Posgrado en Filosofía de la Ciencia - UNAM, Facultad de Estudios Superiores Cuautitlán, Facultad de Música - UNAM.
	13:10 -13:30	Heurística para la generación de preguntas y respuestas a partir de la estructura HTML de documentos	Juan Jesús Sandoval Villanueva, Noé Alejandro Sánchez-Castro, Héctor Jiménez Salazar / CENIDET, UAM- Cuajimalpa
	13:30-13:50	Generación de chatbot para la asistencia en la corrección y	Julio Castañeda Torres / CENIDET
	13:50 -14:05	Preguntas y respuestas	

IX CoLiCo

Coloquio de Lingüística Computacional

Mesas	Horario	Ponencias	Autores
Mesa 3. Lexicografía computacional **Moderador: Gerardo Sierra Martínez**	16:00-16:20	Corpus de concordancias bíblicas	Wendy Anel Vázquez Gómez
	16:20-16:40	Las redes semanticas subyacentes en el Diccionario de Español de México (DEM), El Diccionario de la Lengua Española (DEL) y la Wikipedia en español	Miguel Alejandro Dorantes Cruz / GIL - IINGEN
	16:40-17:00	Experimento de comparación de definiciones del Diccionario del Español de México mediante la aplicación de LSA	Manuel Alejandro Sánchez Fernández, Alfonso Medina Urrea / Centro de Estudios Lingüísticos y Literarios - El Colegio de México.
	17:00-17:20	Las relaciones sintácticas en una tarea de asociación léxica: un acercamiento conexionista	Marco A. Flores-Coronado, Aline Minto-García, Elsa Vargas-García, Ángel. E Tovar, Natalia Arias-Trejo / Laboratorio de Psicolingüística - Facultad de Psicología, UNAM
	17:20-17:35	Preguntas y respuestas	
Mesa 4. Clasificación **Moderadora: Georgina Barraza Carbajal**	17:50-18:10	Modelado de tópicos para la clasificación de novelas de ciencia ficción: una propuesta desde la lingüística computacional	Orlando Gaínza
	18:10-18:30	Detección de tuits noticiosos influyentes mediante aprendizaje automático	Christian E. Maldonado
	18:30-18:50	Caracterización fónica de los cantos de avifauna de la UAM -i utilizando herramientas y metodología de análisis de voz humana	Carlos Misael González Valseca, Irma Urbina Sánchez, Gerardo López Ortega / SUEDIF, CIESAS, Departamento de Biología - UAM-I
	18:50-19:05	Preguntas y respuestas	

PROGRAMA

Miércoles 4 de septiembre de 2019

Mesas	Horario	Ponencias	Autores
Mesa 5 Lenguas indígenas **Moderador: Iván Meza Ruiz**	10:20-10:40	La alineación automática en textos paralelos de lenguas mexicanas: el caso de Corpus Paralelo de Lenguas Mexicanas (CPLM)	Diego Córdova Nieto, Cynthia Araceli Montaño Ramírez / Universidad Autónoma Metropolitana, Universidad Autónoma de Querétaro
	10:40 -11:00	Creación del Subcorpus de Textos Religiosos y Políticos (STRP) del Corpus Paralelo de Lenguas Mexicanas (CPLM)	Margarita Abigail Mota Montoya, Sara Elena Valencia Sánchez, Diana Esperanza Vázquez González / Maestría en Lingüística Aplicada - UNAM, Lengua y Literaturas Hispánicas - UNAM, Gestión y Desarrollo Interculturales - UNAM
	11:00 -11:20	Proyecto de lenguas mexicanas: presentación de la interfaz para el Corpus	Luis Enrique Argota Vega, Emma Laura Rea Silva / Posgrado en Ciencia e Ingeniería de la Computac-ión - UNAM, Licenciatura en lenguas y literaturas Hispánicas - UNAM
	11:20 -11:40	NIERIKA	Vania Ramírez / Licenciatura en Lengua y Literaturas Hispánicas - UNAM,
	11:40 -11:55	Preguntas y respuestas	
Ponencia Magistral		Los recursos lingüísticos actuales en el mundo de la inteligencia artificial	Núria Bel / UPF
Mesa 6 Bases teóricas **Moderador: Axel Hernández Díaz**	12:50 -13:10	Un modelo computacional de la operación Merge del Programa Minimalista	Daniel Martínez-García / Instituto de Investigaciones Filosóficas - UNAM
	13:10 -13:30	Una aplicación de los sistemas de membranas: Análisis sintáctico con autómatas P	Gemma Bel Enguix/ GIL - IINGEN
	13:30 -13:50	Caracterización de los dominios modales epistémico y deóntico de los usos en corpusde las perífrasis deber (de) + infinitivo mediante HAC	Sandra Martín / UNAM
	13:50 -14:05	Preguntas y respuestas	

Mesas	Horario	Ponencias	Autores
Mesa 7 Discurso **Moderadora: Gemma Bel Enguix**	16:00-16:20	La interacción médico-paciente: un análisis pragmático en la consulta de alta	Alejandra Mitzi Castellón Flores, Elisa Orozco Martínez / Facultad de Filosofía y Letras - UNAM
	16:20-16:40	La sexualidad y la cosificación de la mujer en las letras del reggaetón: un enfoque cuantitativo	Andie Álvarez
	16:40-17:00	Patología adjetival: una mirada cualitativa y cuantitativa al uso de los adjetivos en el Corpus Sociolingüístico de la Ciudad de México.	Saúl Hernández / Facultad de Filosofía y Letras - UNAM
	17:00-17:20	Las apps , globalización y el zapoteco variante Istmo: una conexión tecno-lingüística inevitable.	Julissa Valdivieso / Universidad del Valle de México - Campus Tuxtla Gutiérrez
	17:20-17:35	Preguntas y respuestas	
Mesa 8 Clasificación **Moderadora: Fernanda López Escobedo**	17:50-18:10	Clasificación de preguntas en español utilizando una Red Neuronal Convolucional	Alberto Iturbe Herrera, Noé Alejandro Castro-Sánchez, Dante Mújica Vargas / CENIDET
	18:10 -18:30	Clasificación de componentes argumentativos con la Red Neuronal Multilayer Perceptron	Kenia Nieto Benítez, Noé Alejandro Castro Sánchez, Héctor Jiménez Salazar / CENIDET, UAM - Cuajimalpa
	18:30-18:50	Síntesis de voz usando redes neuronales profundas	Emilio Morales, Abel Herrera / Laboratorio de Tecnologías del Lenguaje - Facultad de Ingeniería, UNAM
	18:50-19:05	Preguntas y respuestas	

MESA 1
LINGÜÍSTICA FORENSE

Moderadora: Ana Aguilar Guevara

Valores acústicos de referencia para el análisis forense de grabaciones: el caso de la /a/ en habla leída

10:20-10:40 hrs

Yunuen Sarasuadi Acosta Meza, Iván Vladimir Meza Ruíz, Fernanda López Escobedo / ENAH, IIMAS, LCF - Facultad de Medicina, UNAM.

La voz es única y su estudio es cada vez más requerido en el campo forense, por ello en este trabajo nos concentramos en las propiedades fonéticas de la /a/ en sus sílabas más frecuentes en habla leída. Los datos los obtuvimos del Corpus de Lengua Oral del Español de México (CLOE México, López, F., 2019). Este corpus está integrado por una muestra significativa de grabaciones, se trata de un banco de voces del español de México en habla espontánea y leída en un ambiente controlado. La etiquetación que se realizó fue en más de mil contextos, todos extraídos de este Corpus. Cada etiquetación con tres niveles, que identificaban fonema, etiqueta y palabra fueron procesados a través de herramientas computacionales que permitieron, a través del cálculo de cohefiecientes de probabilidad o Likelihood-ratio, llegar a conocer rasgos de la individualidad de las voces de hablantes de la Ciudad de México. Además, se espera que sea un parámetro útil en la comparación forense de voz de evidencias orales que se utilicen en el contexto jurídico mexicano.

Atribución de autoría mediante aprendizaje supervisado de características estilométricas

10:40- 11:00 hrs

Tonatiuh Hernández García / GIL – IINGEN, UNAM.

Contamos con la transcripción de un texto anónimo titulado ¡El Móndrigo! Bitácora Nacional del Consejo de Huelga y un corpus de obras de autoría comprobada correspondientes a cinco autores probables seleccionados tras una investigación histórica previa del documento. Abordaremos el problema de atribución como una clasificación multiclase. Utilizaremos cinco características estilométricas: por párrafo obtendremos la cantidad de puntos, comas, dos puntos, punto y coma y palabras. A este vector de características se le asignará una etiqueta correspondiente al autor.

Entrenaremos Decision Tree, Support Vector Classifier y Naive Bayes Multinomial para reconocer el estilo de cada autor. Al recibir como entrada los vectores del texto dubitado, los algoritmos serán capaces de asignar cada párrafo al autor con más similitud según el entrenamiento previo. Como salida, los algoritmos entregaran la cantidad de párrafos asignada por autor, lo que permitirá comparar los resultados.

En ejercicios previos, Decision Tree ha obtenido precisión del 0.9754, SVC 0.7950 y NaïveBayes Multinomial: 0.4480. Regularizaremos los datos para optimizar la precisión.

#NiUnaMenos: violencia de género en los microdiscursos de Twitter

11:00-11:20hrs

Michelle Denise Rodríguez Chiw, Tonatiuh Hernández García / Posgrado en Ciencia e Ingeniería de la Computación - UNAM, Licenciatura en Lenguas y Literaturas Hispánicas - UNAM.

Para el presente texto se recopiló un corpus de 2,000 *tweets* con la etiqueta #NiUnaMenos. Con ellos se obtuvieron n-gramas con los cuáles se pretenden encontrar coincidencias de construcciones léxicas para explorar la opinión general respecto a la violencia hacia la mujer. Al realizar un análisis semántico-pragmático de los trigramas obtenidos, se reveló un discurso de violencia intrínsecamente relacionado con los esquemas de poder patriarcales que han sido constantemente señalados por el feminismo y los cuales permiten observar las cuestiones de género, poder y sociedad.

Las frecuencias en los patrones sintácticos son reveladoras: muestran sentimientos de rabia, llamados a reaccionar y el intento por erradicar la violencia, cabe mencionar que sus agresores son principalmente exparejas y parejas. Se muestran los tipos de agresión más frecuentes: empujones y jalones de brazos. El lugar donde las mujeres se sienten más inseguras: el metro. El miedo más frecuente: la muerte.

El estudio utiliza la minería de opinión para llegar a un análisis global del uso de palabras en situaciones precisas; también podemos advertir que el discurso permite observar la función (no solamente lingüística) que le otorgamos a vocablos que presuponen denuncia y que señalan problemas sociales para crear una conciencia social en donde las palabras son un medio de denuncia usual con el que se reitera que los enunciados no pueden deslindarse de su entorno vivo.

Programa BioMetro Fore como una alternativa para el análisis forense

11:20-11:40 hrs

Carlos Misael González Valseca, Sergio Arturo Domínguez Ortega / SUEDIF-CIESAS, ENAH.

La presente ponencia tiene la intención de mostrar el programa BioMet Fore para su uso en el análisis acústico forense, dicho programa fue desarrollado por la Universidad Politécnica de Madrid, el cual ha brindado excelentes resultado en la práctica, lo que ha generado que se utilice en España desde hace algunos años, sin embargo en México es una herramienta poco conocida, aun cuando ésta es capaz de analizar hasta 68 parámetros acústicos de manera automática, pues en palabra de uno de sus creadores "El programa Es una herramienta polivalente que permite la grabación y análisis de la voz, así como su cotejo a tres bandas (modelos, indubitada y dubitada) a partir de las características disfónicas de la misma" (Gómez. 2014) características mismas que están divididas de la siguiente manera:

- Perturbation parameters
- Cepstral Parameters
- Spectral Parameters
- Biomechanical Parameters
- Temporal Parameters
- Glottal GAP Parameters
- Tremor Parameters

Con lo anterior, queremos mostrar por medio de una llamada de extorsión, cuáles fueron los resultados obtenidos, así como también la familiarización de la interfase del programa para que el público se dé cuenta de que es una herramienta relativamente sencilla de usar y una alternativa más para la lingüística forense.

MESA 2
ENSEÑANZA
Moderadora: Helena Gómez Adorno

Asistente computacional para el diseño de instrumentos de evaluación de la comprensión de lectura

12:50-13:10 hrs

Carlos César Jiménez, José Luis Sánchez García / Posgrado en Filosofía de la Ciencia – UNAM, Facultad de Estudios Superiores Cuautitlán y Facultad de Música - UNAM.

Una de las técnicas que suelen utilizarse para evaluar la comprensión que un agente tiene de un texto es señalar si determinados enunciados son consecuencia lógica de dicho texto. La generación de dichos enunciados para conformar un instrumento de evaluación no es trivial y supone por lo menos: (1) un buen análisis lógico-semántico del texto en cuestión, (2) un uso sistemático de varios sistemas lógicos de trasfondo, (3) una teoría de los procesos de comprensión y adquisición del lenguaje.

Nuestra propuesta consiste en asistir el proceso de diseño de instrumentos de evaluación de la comprensión de textos escritos en diversas lenguas (inicialmente alemán, español e inglés) recurriendo a esta técnica. Para ello, además de haber desarrollado una metodología lógico-lingüística robusta, desarrollamos y usamos software para (1) mostrar una gama ordenada de estructuras lógico-semánticas "subyacentes" a los textos usados para evaluar, (2) generar sistemáticamente "reactivos V/F/ND" válidos a partir de los análisis previos.

Entre las herramientas de software que hemos usado se encuentran NLTK, Haskell y CoQ.

Este proyecto parte de trabajo previo de investigación y diseño de evaluaciones para 53 especialidades diferentes realizado en el Centro de Idiomas de la FESC durante varios años.

Heurística para la generación de preguntas y respuestas a partir de la estructura HTML de documentos

13:10- 13:30 hrs

Juan Jesús Sandoval Villanueva,
Noé Alejandro Sánchez-Castro, Héctor Jiménez Salazar
/ CENIDET, UAM- Cuajimalpa.

En este trabajo se describe una heurística implementada en un sistema para realizar evaluaciones automáticas por medio de la generación de preguntas conceptuales y respuestas de opción múltiple. El propósito es evaluar la comprensión de documentos. El método consiste en tres pasos: (1) identificar patrones para el reconocimiento de títulos y subtítulos; (2) identificar el nivel de profundidad de ellos, lo que servirá para identificar las posibles preguntas conceptuales y (3) generar plantillas de preguntas. Las posibles respuestas se extraen de los párrafos asociados a los títulos y subtítulos. Se generarán 4 posibles respuestas, siendo una de ellas la correcta. Las pruebas se realizan con páginas de Wikipedia.

Generación de chatbot para la asistencia en la corrección y análisis de errores ortográficos

13:30-13:50 hrs

Julio Castañeda Torres / CENIDET.

En este trabajo se desarrolló un chatbot utilizando la plataforma Dialogflow, que tiene como finalidad apoyar a los alumnos que presenten problemas de escritura, específicamente faltas o errores de acentuación e intercambio de letras, como lo son "c/s/z" y "b/v". La conversación del chatbot está conformada por cuatro módulos: el primero de ellos tiene la función de responder a dudas acerca del uso de las reglas ortográficas, funcionando como un módulo informativo. El segundo módulo es capaz de proporcionar ejemplos sobre el uso correcto de dichas reglas. El tercer módulo tiene la función de mostrar ejercicios para que el alumno los resuelva y practique su conocimiento, estos ejercicios deben ser contestados mediante las conversaciones. El cuarto módulo, a diferencia del primero, funciona como un consultor respondiendo, en primer lugar, en el cómo se acentúan las palabras, a lo cual el agente responde indicando el tipo de acento de cada palabra (aguda, grave, esdrújula, sobreesdrújula) y en qué silaba lleva el acento ortográfico; y en segundo lugar en el cómo se escribe una palabra, considerando los usos de "c/s/z" o "b/v".

Martes 3 de septiembre del 2019

MESA 3
LEXICOGRAFÍA COMPUTACIONAL

Modera: Gerardo Sierra Martínez

Corpus de concordancias bíblicas

16:00-16:20 hrs

Wendy Anel Vázquez Gómez
Posgrado en Lingüística- UNAM.

En 2013 me doctoré en Lingüística por la UNAM; mi área de estudio es la Lexicografía. Para la realización de mi tesis doctoral elaboré un corpus cuya información se centró en aquella correspondiente a los personajes que son participantes, o bien, referidos en los documentos que constituyen al Nuevo Testamento. Realicé una minería de datos en el conjunto de estos documentos, no solo en la ubicación y señalización de cada participación o referencia de cada uno de los personajes, sino también en la localización de la información contextual ofrecida en el texto bíblico en torno suyo, la cual sirvió para definir sus biografías como producto resultante de la investigación. Asimismo, dada la organización y presentación usual de los documentos bíblicos, donde se vinculan hechos entre textos (en la parte inferior de las páginas aparecen redirecciones a otros documentos donde se mencionan hechos relacionados), una aportación más de mi trabajo fue un corpus de concordancias que incluí en el propio diccionario; realicé un corpus lingüístico con información del Nuevo Testamento con el fin de ubicar a los personajes y a sus biografías, mas también logré una identificación de concordancias en torno suyo, lo que es característico en los textos bíblicos para su comprensión.

Las redes semánticas subyacentes en el Diccionario de Español de México (DEM), el Diccionario de la Lengua Española (DEL) y la Wikipedia en español

16:20- 16:40 hrs

Miguel Alejandro Dorantes Cruz / GIL – IINGEN.

Las redes semánticas son una forma de representar el conocimiento por medio de nodos unidos por aristas. Entendiendo los nodos como palabras y las aristas como relaciones semánticas, la lingüística ha encontrado distintas formas en que las unidades léxicas se relacionan. Entre las relaciones semánticas más comunes se han encontrado las siguientes: hiponimia, hiperonimia, meronimia, holonimia, sinonimia y antonimia, entre otras.

El presente trabajo tiene como objetivo principal presentar la metodología usada para la extracción de las redes semánticas que subyacen en el Diccionario de Español de México (DEM), el Diccionario de la Lengua Española (DLE) y Wikipedia en español.

Adicionalmente, se presentarán los primeros análisis contrastivos de algunas de las redes, pues nos han permitido observar las clases o grupos más generales a las que cada obra lexicográfica categoriza a las palabras del español. Es decir, las formas en que se estructura el significado en estas tres obras lexicográficas.

Se espera que tanto las relaciones semánticas como las redes semánticas aporten directa e indirectamente a áreas como la lingüística, la psicología, la filosofía, la Inteligencia Artificial, entre otras.

Experimento de comparación de definiciones del Diccionario del Español de México mediante la aplicación de LSA.

16:40-17:00 hrs

Manuel Alejandro Sánchez Fernández, Alfonso Medina Urrea
Centro de Estudios Lingüísticos y Literarios -
El Colegio de México.

En este trabajo se presentan los resultados de la aplicación de un análisis de semántica latente (LSA) a partir de las definiciones del Diccionario del Español de México (DEM) con el fin de detectar afinidades entre unidades léxicas. Para ello, se establece una lista preliminar de 75 pares de palabras a modo de prueba que reflejan relaciones semánticas de hiperonimia, hiponimia y sinonimia. Aunque en los análisis lingüísticos se sostiene que la sinonimia no existe, la propuesta es que es posible detectar SINÓNIMOS CERCANOS (*near-synonym*) entendidos como unidades léxicas que pueden utilizarse en un mismo contexto.

Posterior a la aplicación y evaluación de la semejanza entre las definiciones de las palabras del conjunto prueba, se extraen los diez candidatos más probables de ser afines a una unidad léxica dada del DEM a partir de un script secundario. Se propone además una comparación del mismo método utilizando los ejemplos de contexto de uso que el mismo diccionario tiene.

Las relaciones sintácticas en una tarea de asociación léxica: un acercamiento conexionista

17:00- 17:20 hrs

Marco A. Flores-Coronado, Aline Minto-García,
Elsa Vargas-García, Ángel. E Tovar, Natalia Arias-Trejo /
Laboratorio de Psicolingüística - Facultad de Psicología, UNAM.

El léxico mental se encuentra ordenado en redes léxicas (RL) donde la proximidad entre diferentes palabras es determinada por su cercanía semántica. Sin embargo, se ha apuntado que las relaciones gramaticales entre palabras también estructuran las RL; evidencia de ello es que en Tareas de Asociación Léxica (TAL) existe una tendencia a dar respuestas de tipo paradigmático (e.g., perro-gato).

Exploramos si las RL siguen una estructuración semántica y sintáctica mediante una simulación computacional de datos empíricos. En una TAL a adultos mayores (n = 60) encontramos mayor frecuencia de respuestas paradigmáticas que sintagmáticas (e.g., perro-bonito). Para la simulación se construyeron 2 arquitecturas de redes neuronales conexionistas diferentes: una lineal y una recurrente. Las redes fueron entrenadas con una gramática artificial formada por sustantivos, verbos y adjetivos. Se generaron los vectores oracionales según las regularidades semánticas del español. El output constó de un autocodificador que debía replicar el input en 3 diferentes tareas por simulación (identificar información semántica y/o sintáctica). La ejecución de las redes sugiere una insuficiencia de la información puramente semántica para que emerjan patrones de respuesta similares a los datos empíricos de humanos, por lo que se apoya la estructuración semántica y sintáctica de las RL.

Martes 3 de septiembre del 2019

MESA 4
CLASIFICACIÓN

Moderadora: Georgina Barraza Carbajal

Modelado de tópicos para la clasificación de novelas de ciencia ficción: una propuesta desde la lingüística computacional

17:50- 18:10 hrs

Orlando Gaínza.

El uso de técnicas de reducción de características para la representación de documentos en la clasificación automática de textos (multi-word, LSA, LSI, LDA) ha sido un campo fecundo de investigación en los últimos años y tiene importantes aplicaciones en manejo de contenido, minería de opinión, análisis de sentimientos y búsqueda contextual [1]. Dicho enfoque ha sido probado en todo tipo de documentos con buenos resultados: desde reseñas y artículos académicos hasta tweets y comentarios de Facebook [2]. Sin embargo, la clasificación de textos literarios a partir de este método permanece todavía un campo inexplorado (y que vale la pena explorar). El propósito de la presente ponencia es justificar la aplicación de técnicas de reducción de características a un corpus de novelas de ciencia ficción, exponer las particularidades de dicho corpus así como las dificultades que presenta su análisis, y proponer una metodología para la clasificación binaria de novelas de ciencia ficción.

Bibliografía

[1] Para un resumen del estado del arte en la metodología para la clasificación automática de textos, ver: Mita K. Dalai. "Automatic Text Classification: A Technical Review". International Journal of Computer Applications, vol. 28, no. 2. pp. 37-40, 2011.

[2] Para un ejemplo de la aplicación de las técnicas de reducción de características en la clasificación binaria de correos electrónicos, ver: Masoumeh Zareapoor, Seeja K. R, "Feature Extraction or Feature Selection for Text Classification: A Case Study on Phishing Email Detection", IJIEEB, vol.7, no.2, pp.60-65, 2015.

Detección de tuits noticiosos influyentes mediante aprendizaje automático

18:10- 18:30 hrs

Christian E. Maldonado Sifuentes / CIC – IPN.

In the following work, we present an overview of our approach to detect influential news tweets through machine learning algorithms and NLP techniques. It consists of three parts: an in-the-house influence metric, a script which automatically downloads and tags the tweets according to the aforementioned metric, and feature selection and classification of the tweets through machine learning. The features used are a mixture of different types and lengths of n-grams, selected and optimized through supervised and unsupervised filtering algorithms. Several works on social media influence exist since the early years of these online networking platforms, nevertheless, most are based on network graph analysis, and measure the influence of the source, not the content itself. Of those which measure the influence of the content, the subset based on textual analysis of said content utilize manually tagged, massive corpora (millions of items). The main strengths of our approach are to provide an end-to-end workflow with little or no human intervention, while reaching an accuracy in excess of 70% with a small corpus of tweets (as few as 4000), and processing times on the range of seconds -instead of hours or days-, thus, providing an excellent opportunity for the development of business applications.

Keywords: Twitter, News, Influence, Machine Learning, NLP

Caracterización fónica de los cantos de avifauna de la UAM -i utilizando herramientas y metodología de análisis de voz humana.

18:30- 18:50 hrs

Carlos Misael González Valseca, Irma Urbina Sánchez, Gerardo López Ortega / SUEDIF-CIESAS, Departamento de Biología. UAM-I.

El presente trabajo tiene la intención de mostrar cómo, a partir de la metodología utilizada para generar triángulos vocálicos en alguna lengua, generar una propuesta de inventario fonológico, también es posible categorizar los sonidos producidos por diferentes especies de aves. La distribución formántica de estos sonidos permite deducir ciertas características morfológicas del aparato fonador de estas. Todo lo anterior a partir de teorías sobre la forma en que podemos determinar la clasificación articulatoria de los correlatos acústicos, propios del estudio de la fonética.

El objetivo de este trabajo está en utilizar metodologías de base lingüística de la voz humana para obtener clasificaciones fónicas y relacionarlas con significados dentro del comportamiento de las aves.

Este es un primer trabajo experimental, por lo que primero mostraremos los resultados obtenidos en praat a partir del ploteo de los formantes 1, 2, 3, e incluso 4 (este último formante muy poco contemplado en el análisis de la voz humana), el pitch, y la intensidad. Además, se utilizará el programa BioMetroling, software desarrollado para extraer formantes dentro de una emisión para generar un mapa formántico de manera inmediata. Por último, se intentará determinar si los sonidos emitidos por las aves cuentan con armónicos lo suficientemente estables para averiguar si son semejantes a los de una vocal realizada por la una voz humana.

MESA 5
LENGUAS INDÍGENAS

Moderador: Iván Meza Ruiz

La alineación automática en textos paralelos de lenguas mexicanas: el caso de Corpus Paralelo de Lenguas Mexicanas (CPLM)

10:20 - 10:40 hrs

Diego Córdova Nieto, Cynthia Araceli Montaño Ramírez / Universidad Autónoma Metropolitana, Universidad Autónoma de Querétaro.

Para realizar alineación automática de textos resulta adecuado hacer una clasificación de algoritmos de alineamiento oracional que divida a estos en tres grupos: basados en la longitud, en el lexicón y un híbrido de ambos, como lo mencionan Li y Sun (2010). Los algoritmos basados en el lexicón y los híbridos suelen ser implementados en programas robustos porque usan información léxica, lo que conlleva un costo computacional mucho mayor, lo que se traduce en un mayor tiempo de ejecución. Por otro lado, los algoritmos basados en longitud, en comparación con los anteriores, tienen un mejor desempeño computacional, ya que solamente se basan en la longitud de las oraciones (ya sea por cantidad de caracteres o palabras) y no utilizan información léxica, siempre y cuando exista el mínimo ruido posible entre las oraciones que conforman tanto al texto fuente como al texto destino. Con el fin de crear una interfaz digital que permita la visualización de textos en diferentes lenguas mexicanas con alineamiento oracional, así como mejorar la eficiencia del trabajo de alineamiento, hicimos uso de una implementación del algoritmo Gale & Church en código escrito en Python, el cual pertenece al tipo de algoritmos basados en la longitud. Se ha seleccionado este algoritmo, además, porque es independiente de la lengua, es decir, que podemos elegir un texto fuente en cualquier lengua para alinearlo con el texto destino escrito en cualquier otra lengua. Lo cual resulta de suma importancia para el proyecto que llevamos a cabo al trabajar con lenguas de familias muy disímiles cuyas grafías y estructuras morfosintácticas son muy variadas.

Creación del Subcorpus de Textos Religiosos y Políticos (STRP) del Corpus Paralelo de Lenguas Mexicanas (CPLM)

10:40- 11:00 hrs

Mota Montoya, Margarita Abigail, Valencia Sánchez Sara Elena, Vázquez González Diana Esperanza / Maestría en Lingüística Aplicada – UNAM y GIL-IINGEN, Lengua y Literaturas Hispánicas - UNAM, Gestión y Desarrollo Interculturales – UNAM.

El Corpus Paralelo de Lenguas Mexicanas es un proyecto a cargo del Grupo de Ingeniería Lingüística el cual tiene como objetivo poner a disposición textos bilingües en español con diferentes lenguas mexicanas, el subcorpus de textos religiosos y políticos forma parte de dicho proyecto.

Los textos religiosos que conforman este subcorpus son la biblia en 29 lenguas con al menos una de sus variantes. La elección de la biblia se debe a que es de fácil acceso y se encuentra traducido en diferentes lenguas incluyendo las mexicanas, además de diferentes variantes, por lo que su utilidad es innegable para el Procesamiento del Lenguaje Natural.

Las ventajas del empleo de estos textos además de las ya mencionadas son que se encuentran escritos por diversos autores, contiene diversos modos del discurso como los son la narración, descripción y diálogo.

En los textos políticos se incluyó la constitución mexicana en 40 lenguas mexicanas, textos explicativos (en 28 lenguas mexicanas) sobre diferentes temas como la constitución, la nacionalidad, derechos fundamentales, garantías individuales, derechos sociales, territorio, el gobierno, entre otros.

Asimismo, se incluyó la *Declaración Universal de los Derechos Humanos* en 14 lenguas mexicanas (huasteco, maya, mazahua, mazateco, mixe, mixteco, náhuatl, otomí púrepecha, tojolabal, totonaco, tzetzal, tzotzil, zapoteco).

Proyecto de lenguas mexicanas: presentación de la interfaz para el Corpus Paralelo de Lenguas Mexicanas (CPLM)

11:00- 11:20 hrs

Luis Enrique Argota Vega, Emma Laura Rea Silva / Posgrado en Ciencia e Ingeniería de la Computación – UNAM, Licenciatura en lenguas y literaturas Hispánicas – UNAM.

La lingüística de corpus es un enfoque metodológico para el estudio de las lenguas. En la antigüedad era un proceso que se realizaba de forma manual, hoy en día se obtienen mayoritariamente mediante un proceso automatizado. En México existe una amplia variedad de lenguas, muchas de ellas se están extinguiendo, por ello es muy importante preservarlas y evitar su total desaparición. En función de esto, el Grupo de Ingeniería Lingüística de la UNAM desarrolla diversos corpus paralelos entre lenguas mexicanas y español, constituidos por una amplia colección de documentos de diferentes géneros, las lenguas que se trabajan actualmente son: chol, maya, mazateco, mixteco, náhuatl y otomí, además de un extenso catálogo de lenguas que conforman los corpus de textos de la biblia y la constitución. El CPLM pertenece al proyecto "Extracción léxica multilingüe para lenguas de escasos recursos y su aplicación a las lenguas mexicanas", para que el corpus pueda ser consultado se trabaja en la aplicación web para la gestión de corpus paralelo. En este trabajo, se presenta como se ha realizado el corpus paralelo y una propuesta de interfaz para la aplicación web.

Referencias:
Gutiérrez, X., Vilchis E., Cerbón R. (2017) "Recopilación de un corpus paralelo electrónico para una lengua minoritaria: el caso del español-náhuatl". Ciudad de México. Instituto de Ingeniería.
Disponible en: https://www.pcnt.inah.gob.mx/pdf/14290504602.pdf
Mockups, B. (2016). Balsamiq. Consultado el, vol. 1.
Disponible en: https://balsamiq.com/ Sierra G. (2017) Introducción a los corpus lingüísticos.
Ciudad de México, Instituto de Ingeniería, Universidad Nacional Autónoma de México.

NIERIKA

11:20 - 11:40 hrs

Vania Ramírez / Licenciatura en Lengua y Literaturas Hispánicas, Front-End Developer en Citibanamex.

NIERIKA es una red social que pretende colaborar a la preservación de lenguas indígenas mediante la captación de corpus a través de las publicaciones en la red social. Las publicaciones contienen dos inputs de publicación, el primero para publicar en la lengua originaria y el segundo para añadir la traducción al español realizada por el mismo autor, además opcionalmente obtiene datos de la información del perfil del autor como la geolocalización, la variante lingüística a la que pertenece, el tipo de contenido y algunas otras características de la lengua originaria. La red social está construida con React Js, Firebase, es responsiva y está deployada en GitHub Pages. A la data recabada se le aplican procesos de NLP para obtener estadísticas, para ello se ocupa el lenguaje Python y librerías de NLP. Esta Red social ha sido utilizada en una prueba piloto por hablantes de lenguas originarias y se pretenden exponer los resultados obtenidos de la experiencia del usuario y del análisis obtenido de la data procesada. En un segundo alcance está red social será mejorada con el feedback de los usuarios y pretende optimizar la metodología de procesamiento del lenguaje para permitir la búsqueda de contenido por palabra- traducción.

MESA 6
BASES TEÓRICAS

Moderadora: Axel Hernández Díaz

Un modelo computacional de la operación Merge del Programa Minimalista

12:50 -13:10 hrs

Daniel Martínez-García / Instituto de Investigaciones Filosóficas – UNAM.

El Programa Minimalista, la última etapa del generativismo iniciado por Noam Chomsky, propone que la sintaxis de todas las lenguas del mundo está regida por una única operación binaria de ensamblaje: la operación *Merge*.

Dicha operación constituye el centro del presupuesto teórico de la "Gramática universal" la cual se ha reducido debido a nuevos datos en neurología, genética y adquisición del lenguaje. Merge está caracterizada como una operación recursiva que "toma dos objetos sintácticos, los fusiona y devuelve un objeto sintáctico nuevo".

¿Qué tan factible es que dicha operación pueda crear oraciones gramaticales en las lenguas del mundo?, ¿cómo podemos verificar que efectivamente la sintaxis funciona así?

Dicho esto, mostraré un modelo algorítmico el cual simula la operación Merge elaborado en Python el cual toma dos ítems léxicos y los fusiona en un nuevo objeto sintáctico, de manera recursiva, hasta que genera oraciones gramaticales del español. A manera de corpus, y como simulación del lexicón, utilizo 500 tweets por medio de minería de datos. Los ítems léxicos de este corpus se aíslan y se procesan en tiempo real para construir oraciones gramaticales en el español por medio de Merge. Lo que se espera es que, conforme a los postulados teóricos del Programa Minimalista, este modelo computacional pueda elaborar sintácticamente oraciones bien formadas.

Una aplicación de los sistemas de membranas: análisis sintáctico con autómatas P

13:10- 13:30 hrs

Gemma Bel Enguix / GIL – IINGEN, UNAM.

La computación mediante membranas (Paun, 2000) es una rama de la computación natural que investiga modelos de computo basados en la estructura y funcionamiento de los las células vivas y en sus interacciones en tejidos u otras estructuras biológicas. En general, los sistemas de membranas se usan como artefactos generativos, pero poseen una gran flexibilidad. Esto ha permitido diseñar distintas aplicaciones en campos tan diversos como el hardware o la lingüística formal. Este artículo aborda la capacidad reconocedora de dichos sistemas, mediante la definición y aplicación de los llamados autómatas P (Gramatovici, Bel-Enguix, 2003). El trabajo muestra un método de análisis sintáctico del lenguaje natural con membranas activas que produce árboles de dependencias. Para llevar a cabo el parsing se usan un tipo de estructuras llamadas bubble trees (Gladkij, 1968). El mecanismo es capaz de de reconocer estructuras lingüísticas que no son libres de contexto con una complejidad O(n). Adicionalmente, se presenta el desempeño del algoritmo para el análisis de Gramáticas Contextuales.

Referencias:

1. Gramatovici, R. & Bel-Enguix, G. (2005), Parsing with P automata, in Ciobanu, G., Paun, Gh., Pérez-Jiménez, M.J. (eds.), Applications of Membrane Computing, Berlin, Springer-Verlag (Natural Computing Series): 389-410.
2. A.V. Gladkij: On Describing the Syntactic Structure of a Sentence (inRussian). Computational Linguistics, 7, Budapest, 1968.
3. Gh. Paun: Computing with Membranes. Journal of Computer and System Sciences, 61(2000),108-143.

Caracterización de los dominios modales epistémico y deóntico de los usos en corpus de las perífrasis deber (de) + infinitivo mediante HAC

13:30 - 13:50 hrs

Sandra Martín
Universidad Nacional Autónoma de México.

En los últimos años, la modalidad entendida como "la categoría lingüística que se refiere al estatus factual de una proposición" (Palmer 1986:17-18) ha ganado terreno entre los lingüistas, especialmente dentro de la Gramática Cognitiva.

No obstante, al ser un dominio conceptual y no una categoría gramatical (Narrog 2012: 1), la operacionalización de su semántica y su pragmática además de presentar severas dificultades es altamente cuestionable.

Algunos estudios semántico-pragmáticos de corte cognitivo han propuesto analizar algunas palabras funcionales desde su polisemia de un modo relativamente sistemático, especialmente Stefan Gries (2010) con el enfoque de corpus Behavioral Profiles (BP), mismo que se ha retomado para el presente análisis.

El acercamiento de los BP se sustenta en la teoría distribucional, por lo tanto, resulta ideal para caracterizar dominios modales a partir de analizar en un en corpus los patrones contextuales de modalizadores como el auxiliar deber, vinculado con más de una modalidad.

La presente comunicación busca, de manera muy general, caracterizar los dos dominios modales más reconocidos, epistémico y deóntico, mediante la clusterización del análisis en corpus de la perífrasis deber (de) + infinitivo; y, de manera particular, aportar a la minería de textos mediante el refinamiento en el etiquetado de los BP.

Referencias

Bybee, J. L., & Fleischman, S. (Eds.). (1995). Modality in grammar and discourse. Amsterdam; Philadelphia: J. Benjamins.
Cornillie, B., & Izquierdo Alegría, D. (Eds.). (2017). Gramática, semántica y pragmática de la evidencialidad (Primera edición). Pamplona: EUNSA.
Gries, S. T. (2010). Behavioral profiles: A fine-grained and quantitative approach in corpus-based lexical semantics. The Mental Lexicon, 5(3), 323–346.

MESA 7
DISCURSO
Moderadora: Gemma Bel Enguix

La interacción médico-paciente: un análisis pragmático en la consulta de alta

16:00- 16:20 hrs

Alejandra Mitzi Castellón Flores, Elisa Orozco Martínez
Facultad de Filosofía y Letras – UNAM.

La relación entre médico-paciente resulta interesante, pues este tipo de conversación se presenta durante la consulta médica; asimismo, suele ser muy estandarizada. De esta manera, se contó con 21 transcripciones y audios de conversaciones de pacientes dados de alta, así como datos sociodemográficos de los mismos. Con dichos recursos se optó por la realización de un etiquetado pragmático conversacional XML en la plataforma en línea Code Beautify y de un análisis crítico de los actos del habla.

Para estos fines, se emplearon los postulados de Roman Jakobson entorno a las seis funciones del lenguaje, en específico a la función fática; igualmente, los principios de J.L Austin de la Teoría de los enunciados realizativos y de Searle las Condiciones de adecuación de los actos de habla

Esta aportación pretende hacer una reflexión, descripción y propuesta sobre la interacción lingüística entre pacientes y médicos en la consulta médica, a partir de los resultados obtenidos del ya mencionado etiquetado pragmático, así como de los procesos presentes en la conversación como la cortesía, posicionamiento de los hablantes, turnos que permiten examinar desde la pragmática. Se observarán las características conversacionales, deficiencias y aciertos, mediante las que se ofrecerá una proposición conversacional para la consulta médica.

La sexualidad y la cosificación de la mujer en las letras del reggaetón: un enfoque cuantitativo

16:20- 16:40 hrs

Andie Álvarez
Universidad Autónoma de Querétaro.

Estudios de corte cualitativo han puesto de manifiesto que el reggaetón es un género musical polémico debido a su contenido explícito de naturaleza sexual y misógina. Para nuestra investigación se realizó un estudio a partir de un corpus extenso de letras de canciones de cantantes y autores reggaetoneros, tanto hombres como mujeres. Primero, se estableció una comparación de dicho corpus de reggaetón con el CREA, de la que se obtuvo un listado de las palabras clave. Después, se realizó una comparación entre las letras de las canciones interpretadas por los cantantes y las letras de las cantantes mediante la extracción de palabras clave y la generación de perfiles léxicos. Se analizaron concordancias, colocaciones y diversos patrones léxico-gramaticales. Los resultados muestran diferencias y similitudes entre el contenido de las canciones de los y las cantantes. Por ejemplo, se analizaron las diferencias entre el uso de las palabras que hacen referencia tanto al cuerpo femenino como al masculino, el uso de los pronombres clíticos de segunda persona singular en estructuras siempre seguidas por verbos de movimiento y modificación, asimismo descubrimos que con frecuencia utilizan los clíticos de primera persona singular seguidos por verbos de percepción dándonos una idea de mansplaining.

Patología adjetival: una mirada cualitativa y cuantitativa al uso de los adjetivos en el Corpus Sociolingüístico de la Ciudad de México.

16:40- 17:00

Saúl Hernández
Facultad de Filosofía y Letras – UNAM.

El presente proyecto realiza un estudio sobre el Corpus Sociolingüístico de la Ciudad de México, ya que en dicha investigación se clasifican a los hablantes según tres variantes sociolingüísticas (edad, género y nivel sociocultural). En este trabajo se parte desde la proposición que el habla de cada estrato posee cualidades y diferencias ante las otras para poder ser considerado como dialecto dentro de la lengua estándar. A pesar de lo evidente que pueda parecer esta afirmación, no hay muchos estudios de carácter cuantitativo y cualitativo sobre los sociolectos que se interrelacionan en una región específica.
Por lo tanto, el objetivo es realizar un acercamiento a las disimilitudes que existen del habla entre cada estrato, haciendo un análisis de los adjetivos empleados en las entrevistas que integran el corpus. Para llegar a esta meta, primero se exploró, de forma estadística, las características numéricas de cada dialecto; en una segunda instancia, se expresaron las realidades de uso a través de un conteo de N-gramas y de la clasificación de funciones morfosintácticas. Los resultados arrojaron datos que marcan grados de independencia y, por consiguiente, de diferenciación entre los dialectos de un idioma estándar, y, dentro de éstos, entre cada variante sociolingüística estudiada.

Las apps, globalización y el zapoteco variante Istmo: una conexión tecno-lingüística inevitable.

17:00- 17:20 hrs

Julissa Valdivieso
Universidad del Valle de México - Campus Tuxtla Gutiérrez.

Juchitán o Ixtaxochitlán (lugar de las flores blancas) es una ciudad zapoteca ubicada al sureste del estado de Oaxaca, en la región del Istmo de Tehuantepec. Ahí prevalece la lengua originaria de los binni zaa, el zapoteco. De acuerdo al INEGI, 53 mil habitantes de un total de 100 mil son bilingües pues hablan español L2 y diidxa zaa L1 (zapoteco).

Una prestación lingüística ocurre cuando una palabra utilizada en otro idioma; en Juchitán, de acuerdo a la clasificación de préstamos léxicos, se da el tipo extranjerismo.

Una de las causas principales es la globalización y proliferación de productos que facilitan la comunicación o la vida, aunque esto vas más allá de los términos que no tienen una traducción al zapoteco o al español, pues los hablantes del zapoteco se apropian de esos extranjerismos y le dan la ligera variación fonética, nace una variante llamada Zapotenglish.

Por ejemplo, una tarde mientras conversas en zapoteco con tu madre, ella puede mencionar Facebook, whatsapp, laptop, internet, wifi, tablet, pues son nombres comunes y algunos propios esto sin dejar de lado su lengua madre, ni la esencia poética del zapoteco.

Hay quienes aún estigmatizan las prestaciones lingüísticas y las tachan de aberrantes e impuras, pero son los hablantes quienes hacen caminar las lenguas e idiomas y que, para el nacimiento de otras variantes, idiomas, lenguas desde siempre se han recurrido a prestaciones lingüísticas.

Referencias

INEGI Instituto Nacional de Estadística y Geografía (2013)
John Lyons Introducción al lenguaje y la lingüística Traducc. Ramón Cerdá Edit. Teide Barcelona,1984.
Jose Luis Avila La era Neoliberal Edit. Oceano 2006

MESA 8
CLASIFICACIÓN

Moderadora: Fernanda López Escobedo

Clasificación de preguntas en español utilizando una Red Neuronal Convolucional

17:50- 18:10 hrs

Alberto Iturbe Herrera, Noé Alejandro Castro-Sánchez, Dante Mújica Vargas / CENIDET.

En los últimos años, las técnicas convencionales de Procesamiento de lenguaje natural se han visto igualadas e incluso sobrepasadas por los modelos de Aprendizaje Profundo (*Deep Learning*). Estos grandes avances se deben a la rigidez de estas técnicas, en cambio, los modelos de Inteligencia artificial de última generación, como las redes neuronales artificiales que permiten solucionar distintas tareas del Procesamiento de lenguaje natural con una gran efectividad, como: clasificación de textos, traducción automática, pregunta-respuesta, etiquetado PoS, generación de subtítulos, análisis de sentimientos, entre otras. Enfocándose en la tarea de pregunta-respuesta se han utilizado distintos paradigmas para resolver tal problema, desde reconocimiento de patrones, Procesamiento de lenguaje natural, enfoques probabilísticos, Inteligencia artificial, entre otras. Sin embargo, los distintos trabajos en esta área se enfocan en el idioma inglés. En este trabajo se presenta una red neuronal convolucional que realiza aprendizaje supervisado para generar un modelo capaz de clasificar preguntas en español, arrojando como salida información que puede ser utilizada para aportar más información en tareas como preguntas-respuesta, en la cual, ayudaría a la recuperación de información relacionada de esa forma evitar la selección de respuestas candidatas que no se encuentren estrechamente relacionadas al foco de la pregunta.

Clasificación de componentes argumentativos con la Red Neuronal Multilayer Perceptron

18:10- 18:30 hrs

Kenia Nieto Benitez,
Noé Alejandro Castro Sánchez, Héctor Jiménez Salazar
CENIDET, UAM – Cuajimalpa.

El incremento de la información en la web (periódicos en línea, reseñas de productos, blogs, entre otros) y la creación de redes sociales proporcionan abundante información, en la cual se encuentran y analizan argumentos generados por algún usuario. Existen diferentes enfoques para evaluar los argumentos, tales como retóricos, que se enfoquen en la audiencia y la intensión persuasiva; los dialógicos, que describen las formas en que los argumentos son conectados en las estructuras del dialogo entre dos personas; y los monológicos, que se enfocan en identificar la estructura y las relaciones entre los componentes (premisas y conclusiones) de un argumento dado. Para el análisis de argumentos, además de los enfoques lingüísticos, se pueden emplear métodos computacionales, como Aprendizaje automático (Machine Learning). Entre los más utilizados se encuentran Support Vector Machine y Naive Bayes. Sin embargo, estos modelos pueden ser simulados por una Red Neuronal Artificial. El presente trabajo se enfoca analizar argumentos desde un punto de vista monológico, proponiendo el uso de una Red Neuronal de tipo Multilayer Perceptron (MLP) junto con marcadores del discurso para identificar los componentes de argumentos analizando textos de un corpus en inglés.

Síntesis de Voz usando Redes Neuronales Profundas

18:30- 18:50 hrs

Emilio Morales, Abel Herrera
Laboratorio de Tecnologías del Lenguaje - Facultad de Ingeniería, UNAM.

La síntesis de voz es una tarea imprescindible en la interacción usuario-maquina. El objetivo de obtener voz sintética indistinguible de las grabaciones humanas es aún válido. En los últimos años, las técnicas de Redes Neuronales Profundas han obtenido importantes resultados en síntesis de voz. Dentro de las redes neuronales profundas, los modelos generativos se proponen simplificar el proceso de síntesis de voz tradicional mediante la sustitución de la producción de características lingüísticas y acústicas con una sola red neuronal. Los modelos antes mencionados se analizaron a través del sintetizador Tacotron 2. Este es un modelo generativo completo (end-to-end). En este trabajo se exploran las capacidades de esta técnica para la síntesis del español hablado en México

Notas

IXCoLiCo

Coloquio de Lingüística Computacional

Dr. Enrique Graue Wiechers
Rector

Dr. Leonardo Lomelí Vanegas
Secretario General

Ing. Leopoldo Silva Gutiérrez
Secretario Administrativo

Dr. Alberto Ken Oyama Nakagawa
Secretario de Desarrollo Institucional

Dr. César Iván Astudillo Reyes
Secretario de Atención a la Comunidad Universitaria

Dra. Mónica González Contró
Abogada General

Mtro. Néstor Martínez Cristo
Directora General de Comunicación Social

Dr. Germán Enrique Fajardo Dolci
Director de la Facultad de Medicina

Dr. Jorge Enrique Linares Salgado
Director de la Facultad de Filosofía y Letras

Dr. Luis A. Álvarez Icaza Longoria
Director del Instituto de Ingeniería

Comité organizador:

Dra. Fernanda López Escobedo
Licenciatura de Ciencias Forenses,
Facultad de Medicina UNAM

Dr. Gerardo Sierra Martínez
Grupo de Ingeniería Lingüística,
Instituto de Ingeniería, UNAM

Con apoyo de PAPIIT IA401419 Y IG400119
y CONACYT FC2016/2225

www.colico.unam.mx