**Universidad del Cauca**

**Instituto de postgrados en Ingeniería Electrónica y Telecomunicaciones**

**Programas de Maestría y Doctorado en Ingeniería Telemática**

**Seminario de Investigación**

***Titulo de la relatoría:***

**ALGORITMO METAHEURÍSTICO PARA LA CONSTRUCCIÓN DE UN IDENTIFICADOR DE PARTES DEL DISCURSO (PART-OF-SPEECH TAGGING)**

**Relator: Msc. Luz Marina Sierra Martínez estudiante de Doctorado**

**Co-relator: PhD. Juan Carlos Corrales**

**Protocolante: Msc. Maritza Mera, estudiante de Doctorado**

**Fecha:** 6 de mayo de 2016

**Hora de Inicio:** 10:15 a.m.

**Hora de Finalización:** 11:15 a.m.

**Lugar:** Universidad del Cauca, Popayán

**Asistentes:**

PhD. Juan Carlos Corrales (Coordinador del seminario)

MSc. Luz Marina Sierra (Relator)

Estudiantes de Maestría y Doctorado en Telemática (U. del Cauca)

Estudiantes de Pregrado de la FIET

**Orden del día:**

1. Presentación a cargo del relator.
2. Intervención del co-relator.
3. Discusión.

**Desarrollo:**

1. **Presentación a cargo del relator:**

La Msc Luz Marina Sierra inicia la presentación de su propuesta de investigación de doctorado, mencionando los temas que abarcará durante la misma, los temas definidos son los siguientes: Contexto del proyecto, Planteamiento del Problema y Objetivos. A continuación se explica en detalle cada tema expuesto durante la presentación.

**Contexto del proyecto**

La Msc Sierra presenta los siguientes conceptos relacionados al desarrollo de su proyecto:

**Identificador de Partes del discurso (POS Tagging)**

POS Tagging es el proceso de marcar palabras en un texto (corpus) basándose en su relación con las otras palabras que se encuentran alrededor (Baeza-Yates and Ribeiro-Neto, 1999). Provee información tanto de la morfología (estructura de las palabras) y la sintaxis (estructura de las oraciones).

Teniendo en cuenta la implementación de un identificador, la Msc Sierra presenta los principales algoritmos de clasificación, en los cuales se destacan los siguientes algoritmos:

* Supervisados:
  + Basados en reglas (Forsati & Shamsfard, 2015).
  + Redes Neuronales (Baeza-Yates and Ribeiro-Neto, 1999).
  + Basados en métodos estadísticos Silva, Silva, & Rodrigues, 2014).
* Semisupervisados / No Supervisados
  + Otros enfoques: Híbridos, metaheurísticos

La Msc Sierra presenta los campos de aplicaciones de los sistemas POST (Baeza-Yates and Ribeiro-Neto, 1999), los cuales se mencionan a continuación:

* Sistemas de reconocimiento de voz
* Conversión de texto a voz
* Traducción automática
* Clasificación de textos
* Extracción automática de información
* Recuperación de información multimedia
* Análisis de sentimientos,
* Resolución de ambigüedades en el significado de las palabras en un contexto,

**Metaheurísticas**

Un heurístico es un “procedimiento simple, a menudo basado en el sentido común, que se supone que ofrecerá una buena solución (aunque no necesariamente la óptima) a problemas difíciles, de un modo fácil y rápido”. (Zanakins y Evans, 1981).

La aplicación de las técnicas Metaheurísticas:

* problemas de optimización: discretos, continuos, binarios.
* problemas que no tienen un algoritmo o heurística específica que dé una solución satisfactoria; o bien cuando no es posible implementar ese método óptimo.

Seguidamente, la Msc presenta las ventajas y desventajas de este tipo de algoritmos, las cuales se describen a continuación:

**Ventajas**:

* Algoritmos de propósito general
* Gran éxito en la práctica
* Fácilmente implementables
* Fácilmente paralelizables
* Robustos y eficientes

**Desventajas**

* Son algoritmos aproximados, no exactos
* Son altamente no determinísticos (probabilísticos)
* Presentan poca base teórica

**Algoritmos Evolutivos**

Están basados en la aplicación de los principios de la biología evolutiva en la ciencia de la computación. En su implementación emplean métodos derivados de la herencia, la mutación, la selección natural y la recombinación. Dentro los principales algoritmos que se destacan bajo este enfoque son: Global Best Harmony Search (Forsati & Shamsfard, 2015) (GBHS) (Omran & Mahdavi, 2008), DEKmeans (Sierra, Cobos, Corrales, 2014) y Algoritmo del pescador (Machado, et.al, 2012 y 2014).

Una vez la Msc finaliza la presentación de los conceptos relevantes para el desarrollo de su investigación, continua con la presentación de la motivación.

**Motivación**

La Msc Sierra menciona el crecimiento de las comunicaciones a través de dispositivos y la necesidad de hacer Procesamiento de Lenguaje Natural para aprovechar dicho crecimiento. Teniendo en cuenta esto, se menciona que el Procesamiento de Lenguaje Natural se puede hacer empleando Identificadores de Partes del Discurso a través de Enfoques Automáticos. En este sentido, se requieren datos de alta calidad para hacer el entrenamiento de los algoritmos.

**Definición del Problema**

La Msc Sierra continúa con la presentación del problema de su investigación. Inicialmente menciona la dificultad presente en el proceso de etiquetado debido a que existe ambigüedad en el uso de la palabra. Adicionalmente, se menciona la importancia de los costos computacionales y la necesidad de tener datos de calidad cuando se emplea un enfoque supervisado.

Considerando lo anterior, la Msc Sierra plantea la siguiente **Hipótesis**.

Es posible proponer un algoritmo (s) basado en Metaheurísticas para encontrar la etiqueta más probable para una palabra dentro de una oración específica para tres lenguas tradicionales.

La hipótesis planteada permite formular la siguiente pregunta de investigación: ***Es posible proponer un algoritmo (s) basado en Metaheurísticas para encontrar la etiqueta más probable para una palabra dentro de una oración específica(contexto) para tres lenguas tradicionales***.

La Msc justifica la pregunta de investigación considerando que los enfoques no supervisados para el problema de etiquetado permiten obtener mejores probabilidades en la búsqueda en tiempos razonables (Forsati & Shamsfard, 2015), resultados comparados con otras aproximaciones y mejorar la robustez de los algoritmos típicamente usados en el área (Araujo, 2007).

**Estado del Arte**

La Msc Sierra presenta los principales trabajos asociados al POS Tagging con el fin de identificar las principales contribuciones de estos a trabajos a su propuesta. Dichas contribuciones se presentan a continuación:

**POS Tagging**

* Comprensión de los aspectos relevantes en un POST
* Diferentes propuestas para construir POS Tagging para lenguas diferentes a las tradicionales.
* La experimentación y los valores esperados en las medidas de desempeño.
* Métodos de estereotipos que categorizan los roles de las palabras.
* Diferentes aplicaciones de los POS Tagging.
* Pasos a tener en cuenta para la construcción de un POS Tagging
* Aplicaciones (diccionarios, requisitos de software, entre otros)

**POS Tagging basados en reglas**

* Combinación de diferentes técnicas para la construcción POS Tagging.
* Estrategias para mejorar su desempeño.
* Descripción de los pasos que ejecuta un POS T
* Conocimiento de los diferentes comportamientos de las palabras en un corpus

**POS Tagging basados en métodos estadísticos**

* Tamaño del corpus y características
* Aportes de los HMM
* Características morfo sintácticas que se deben tener en cuenta en un idioma.

**POS Tagging utilizan algorítmos metaheurísticos**

* Cómo hacer líneas bases
* Cómo combinar estrategias y métodos para desarrollar un POS Tagging.

Considerando lo anterior, la Msc Sierra menciona las principales brechas de investigación identificadas.

* Abordar el problema de etiquetado desde una perspectiva semisupervisada o supervisada con miras a que sea multilenguaje
* No se ha utilizado Metaheurísticas (GBHS, DEKmeans y algoritmo del Pescador) para el problema de etiquetado
* Hay un gran campo de Trabajo en esta área, los trabajos encontrados son recientes

**Objetivos**

De acuerdo a las brechas de investigación identificadas la Msc Sierra plantea los siguientes objetivos.

**General**

Proponer un algoritmo competitivo con el estado del arte, basado en Metaheurísticas para encontrar la etiqueta más probable para una palabra dentro de una oración específica (contexto) para tres lenguas occidentales (Inglés, Castellano y Portugués)

**Específicos**

* Adaptar los algoritmos metaheurísticos Mejor Búsqueda Armónica Global, Evolución Diferencial con K-means y Procedimiento de Búsqueda del Pescador en busca de determinar el que mejores resultados brinde al problema de etiquetado.
* Proponer un algoritmo basado en las metaheurística Mejor Búsqueda Armónica Global, Evolución Diferencial con K-means o Procedimiento de Búsqueda del Pescador para la tarea de identificación de partes del discurso para tres lenguas occidentales.
* Evaluar el nivel de desempeño del identificador de partes del discurso construido considerando aspectos de precisión y recuerdo, mediante el diseño de una prueba aplicada a corpus lingüísticos.

**Discusión**

El Dr Juan Carlos Corrales inicia la co-relatoria, y hace algunas observaciones sobre el cambio de los objetivos. Debido a que no está de acuerdo con cambiar a realizar la investigación con un enfoque multilengual.

A continuación se inicia la discusión.

La relatora argumenta que se basó en la revisión de la literatura y el conocimiento de expertos para cambiar el campo de aplicación debido a que el problema continuará siendo el mismo. Adicionalmente, justifica que se ha realizado un avance en la elaboración del corpus en Nasa-Yuwe, sin embargo, el cambio se hace debido a que la validación requiere un corpus ampliamente probado.

Se cuestiona por qué se cambia el escenario si ya se habían hecho algunos avances empleando Nasa-Yuwe en la construcción de un corpus.

Respuesta: La relatora menciona algunas particularidades especiales del lenguaje en cuanto a la conformación de las palabras y las frases, lo cual ha dificultado la conformación del corpus.

Se le hace la recomendación de continuar con los mismos objetivos de la presentación anterior y para aumentar el impacto de la evaluación, evaluar los algoritmos con el corpus en Nasa-Yuwe y uno de otra lengua.

La relatora comenta los avances realizados sobre el corpus en Nasa-Yuwe y la manera cómo se planea la nueva evaluación empleando corpus de 3 lenguas occidentales diferentes.

Se pregunta sobre el segundo objetivo.

La relatora menciona que se propone adaptar los algoritmos del primer objetivo con el fin de lograr una combinación que permita proponer un nuevo algoritmo en el objetivo dos.

Se le recomienda a la relatora no cambiar de campo de aplicación y continuar con el Nasa-Yuwe pero cambiar la manera como se justifica la evaluación. Debido a que en la literatura será el primer trabajo empleando esta lengua. Adicionalmente, se le propone hacer dos casos de estudio; el primero con Nasa-Yuwe y el segundo con un lenguaje que sea con una estructura diferente, con el fin de evaluar el comportamiento del algoritmo propuesto.

**Referencias**

Alejo Machado, O., Fernández Luna, J., Huete Guadix, J., & Concepción Morales, E. (2012). Fisherman Search Procedure. En Advances in Artificial Intelligence – IBERAMIA 2012 Lecture Notes in Computer Science (Vol. 7637, págs. 291-299). Springer Berlin Heidelberg.

Alejo-Machado, O., Fernández-Luna, J., Huete, J., & Concepción Morales, E. (2014). Fisherman search procedure. Progress in Artificial Intelligence, 2(4), 193-203.

AlSuhaibani, R., Newman, C., Collard, M., & Maletic, J. (2015). Heuristic-Based Part-of-Speech Tagging of Source Code Identifiers and Comments. 2015 IEEE 5th Workshop on Mining Unstructured Data (MUD) (págs. 1-6). Bremen: IEEE. doi:10.1109/MUD.2015.7327960

Araujo, L. (2007). How evolutionary algorithms are applied to statistical natural language processing. Artificial Intelligence Review, 28(4), 275-303.

Ariaratnam, I., Weerasinghe , A., & Liyanage, C. (2014). A shallow parser for Tamil. 2014 International Conference on Advances in ICT for Emerging Regions (ICTer) (págs. 197 - 203). Colombo: IEEE. doi:10.1109/ICTER.2014.7083901

Aziz T, A., & Sunitha, C. (2015). A Hybrid Parts Of Speech Tagger for Malayalam. 2015 International Conference on Advances in Computing, Communications and Informatics (ICACCI) (págs. 1502 - 1507). Kochi: IEEE. doi:10.1109/ICACCI.2015.7275825

Brill, E. (1995). Transformation-based error-driven learning and natural language processing:. Comput. Linguist., 21, 543–565.

Brownlee, J. (2011). Clever Algorithms Nature-Inspired Programming Recipes. Melbourne: lulu.com.

Cobos, C., Estupiñán, D., & Pérez, J. (2011). GHS + LEM: Global-best Harmony Search using learnable evolution models. En Applied Mathematics and Computation (Vol. 218, págs. 2558-2578).

Crețulescu, R., David, A., Morariu, D., & Vințan, L. (2014). Part of Speech Tagging with Naïve Bayes Methods. Proceedings of the 18th International Conference on System Theory Control and Computing (págs. 446 - 451). Sinaia, Romania: IEEE.

CRIC y el Programa de Dllo Rural en la Región de Tierra Dentro Cxhab Wala -PT/CW. (2005). Diccionario Nasa Yuwe - Castellano (Primera ed.). Popayán: Litografía San José.

Dinakaramani, A., Rashel, F., Luthfi, A., & Manurung, R. (2014). Designing an Indonesian Part of speech Tagset and Manually Tagged Indonesian Corpus. 2014 International Conference on Asian Language Processing (IALP) (págs. 66 - 69). Kuching: IEEE.

Farfán Martínez, M., & Rojas Curieux, T. (2010). Zuy Luuçxkwe kwe'kwe’sx ipx kwetuy piyaaka. Cartilla de aprendizaje de nasa yuwe como segunda lengua. Buenos Aires.

Fatwanto, A. (2013). Natural Language Requirements Specification Analysis Using Part-of-Speech Tagging. 2013 Second International Conference on Future Generation Communication Technology (FGCT) (págs. 98 - 102). London: IEEE. doi:10.1109/FGCT.2013.6767215

Forsati, R., & Shamsfard, M. (2012). Cooperation of Evolutionary and Statistical statistical PoS-tagging. 2012 16th CSI International Symposium on Artificial Intelligence and Signal Processing (AISP) (págs. 446 - 451). Shiraz, Fars: IEEE.

Forsati, R., & Shamsfard, M. (2015). Novel harmony search-based algorithms for part-of-speech tagging. Knowledge and Information Systems, 42(3), 709-736.

Forsati, R., Shamsfard, M., & Mojtahedpour, P. (2010). An Efficient Meta Heuristic Algorithm for POS-Tagging. 2010 Fifth International Multi-Conference on Computing in the Global Information Technology (ICCGI) (págs. 93- 98). Valencia: IEEE.

Francis, M., & Ramachandran Nair, K. (2014). Hybrid Part of Speech Tagger for Malayalam. ICACCI, 2014 International Conference on Advances in Computing, Communications and Informatics (págs. 1744 - 1750). New Delhi: IEEE. doi:10.1109/ICACCI.2014.6968565

Gramaticas.net. (2011). Recuperado el 23 de Enero de 2014, de http://www.gramaticas.net/2011

Gupta, S., Malik, S., Pollock, L., & Vijay-Shanker, K. (2013). Part-of-speech tagging of program identifiers for improved text-based software engineering tools. 2013 IEEE 21st International Conference on Program Comprehension (ICPC) (págs. 3 - 12). San Francisco: IEEE. doi:10.1109/ICPC.2013.6613828

Ibarra Quiroga, A. P., Mosquera Ramirez, J. C., & Zuñiga Muñoz, R. F. (2004). Proyecto Apoyo Multimedial Indigena "AMI" Propuesta Metodologica para la Construccion de Software Etnoeducativo. (U. C. Colombia, Ed.) Popayán: Universidad Cooperativa de Colombia.

Ismail, S., Rahman, M., & Al Mumin, M. (2014). Developing an Automated Bangla Parts Of Speech. 2013 16th International Conference on Computer and Information Technology (ICCIT) (págs. 355 - 359). Khulna: IEEE. doi:10.1109/ICCITechn.2014.6997347