Construcción de un corpus de noticias hondureñas para análisis lingüístico por medio de técnicas de procesamiento de lenguaje natural

Jairo Martínez Hernández
UNAH-Tec, Danlí, Honduras

Lester Yonabel Brográn
UNAH-Tec, Danlí, Honduras

Datos de contacto:
Jairo Jonathán Martínez Hernández
UNAH-Tec Danlí
Informática Administrativa
Frente a Hospital Regional de Oriente
Cel: +50499440139
e-mail: jairomart@hotmail.com

Área Temática: Gestión de conocimiento y herramientas inteligentes.

Palabras clave: lenguaje, natural, lingüística, computacional, corpus

Aplicación a Premios: ¿Desea que su investigación en prospecto pueda, si es aceptada, ser considerada para ser premiada como uno de los mejores trabajos de investigación?.

- 1: Sí _____ No _____
- 2: Sí _____ No _____
- 3: Sí _____ No _____
- 4: Sí _____ No _____
- 5: Sí _X___ No _____

Aplicación a Edición especial: Su documento de investigación, de ser aceptado, ¿le gustaría fuera considerado para una Edición especial del Congreso (Por favor ver la información acerca de las Ediciones especiales de la Conferencia en la página web del congreso)? Para ello, es obligatorio elegir y marcar una de las siguientes opciones:

- Revista E&A : Sí __X__ No_____
- Revista CEAT: Sí __ _____ No_____

Lenguaje de presentación durante el Congreso:

¿Inglés?: ___X_____ ¿Español? : _____

Construcción de un corpus de noticias hondureñas para análisis lingüístico por medio de técnicas de procesamiento de lenguaje natural

Jairo Martínez Hernández
UNAH-Tec, Danlí, Honduras

Lester Yonabel Brográn
UNAH-Tec, Danlí, Honduras

Datos de contacto:
Jairo Jonathán Martínez Hernández
UNAH-Tec Danlí
Informática Administrativa
Frente a Hospital Regional de Oriente
Cel: +50499440139
e-mail: jairomart@hotmail.com

Resumen
La necesidad de interactuar de manera más fluída con las computadoras, celulares y otros dispositivos ha llevado a estudiar el lenguaje humano por métodos computacionales. A esta área se la denominado procesamiento de lenguaje natural. El análisis del lenguaje natural es un reto para la lingüística computacional. La forma en la que los humanos procesamos el lenguaje dificulta su análisis. Por ejemplo, una misma palabra puede tener diferentes significados dependiendo del contexto. También, a menudo no se respetan las reglas básicas de redacción y ortografía, es decir, tendemos a cometer errores léxicos y gramaticales que finalmente no imposibilitan la comunicación. El estudio del lenguaje natural comprende la recuperación de información, el análisis de sentimiento, responder a preguntas, la creacción automática del resumen, entre otras tareas. Para la investigación en esta área se requiere de colecciones de prueba que permitan realizar la experimentación. En este artículo se presenta la construcción y composición de una colección de prueba para el procesamiento de lenguaje natural con técnicas computacionales, tomando como base noticias publicas en medios de comunicación digital de Honduras.

Palabras clave: Lenguaje, Corpus, Procesamiento de lenguaje natural

1. Introducción

Los últimos años han marcado una creciente producción de información en forma de texto. La popularización de las páginas web, los blogs, las *wikis*, los foros de discusión y las redes sociales han provocado un aumento significativo de los datos disponibles. Sin embargo, estos datos no son claramente estructurados como los que vienen de sensores automáticos o bases de datos estructuradas, si no que son presentados en lenguaje natural. El lenguaje natural es el producido por humanos con propósitos de comunicación oral o escrita.

El análisis del lenguaje natural es un reto para la computación lingüística. La forma en la que los humanos procesamos el lenguaje dificulta la programación computacional. Por ejemplo, una misma palabra puede tener diferentes significados dependiendo del contexto. También a menudo, no se respetan las reglas básicas de redacción y ortografía, es decir, tendemos a cometer errores léxicos y gramaticales que, sin embargo, no imposibilitan la comunicación. Además, usamos algunas figuras

idiomáticas, expresiones o incluso la ironía para transmitir un mensaje diferente al aparente.

Por eso, la interacción en lenguaje natural con una computadora es todavía un campo de investigación abierto. Las principales empresas en tecnología como Google, Apple y Microsoft han desarrollado asistentes capaces de entender el lenguaje natural, pero a pesar de sus avances, resulta evidente que aún les falta mucho para tener una interacción completa y certera con un humano.

El procesamiento de lenguaje natural puede dividirse en tareas más específicas, entre estas se pueden mencionar la recuperación de información, el análisis de sentimiento, traducción automática y respuesta a pregunta, entre otras. La investigación en las diferentes tareas del procesamiento de lenguaje natural requieren colecciones de documentos creados por humanos sobre los cuales aplicar la experimentación. En este artículo se busca: presentar los conceptos básicos y las tareas relacionadas con el procesamiento del lenguaje natural, describir la contrucción de un corpus de documentos en lenguaje natural a partir de noticias publicadas en medios de comunicación digitales de Honduras, mostrar las características principales de los documentos que conforman el corpus construído.

El artículo se organiza de la siguiente forma: Primero se presentan algunos conceptos generales sobre el procesamiento del lenguaje natural, su definicion, sus niveles y las tareas que comprende. Luego se presenta la necesidad de una colección para la investigación y enseñanza de las diversas técnicas utilizadas en cada una de las tareas y finalmente, se describe la colección resultante.

2. Procesamiento de lenguaje natural

El procesamiento de lenguaje natural (NLP, por sus siglas en inglés) es un área de estudio que busca la forma en que las computadoras pueden ser usadas para procesar el texto en lenguaje natural (Chowdhury, 2003) y la ciencia que estudia el NLP es la lingüística computacional (Gelbukh, 2010). El NLP busca técnicas computacionales para analizar y representar el texto que luego será analizado a uno o más niveles lingüisticos tratando de procesar el lenguaje de forma similar a como lo hace el humano (Liddy, 2001).

Según Liddy (2001) un sistema de NLP ideal debería poder:

- Parafrasear un texto de entrada
- Traducir un texto de un idioma a otro
- Responder preguntas con base en el contenido del texto
- Realizar inferencias sobre el contenido del texto.

Los mayores avances se han realizado en los primeros tres. Los sistemas actuales no son capaces de realizar inferencias y es un campo de investigación activo en la lingüística computacional.

Tradicionalmente se ha dividido el NLP en el procesamiento de lenguaje escrito y oral. El procesamiento del lenguaje oral también se apoya en otras ciencias como la acústica y la fonología. El análisis del discurso oral se centra en la forma en que los sonidos pueden transcribirse a morfemas y palabras. Teniendo en cuenta ambas divisiones, según Liddy (2001) pueden mencionar algunos niveles en el NLP que se presentan en sección 3.

3. Niveles del procesamiento de lenguaje natural

Según Liddy (2001) el NLP puede estudiarse con base en los niveles de análisis língüistico aplicado. Los principales niveles son:

Análisis fonológico. Este nivel se encarga del estudio de los sonidos que ocurren dentro y durante la pronunciación de las palabras. Este análisis se vincula de cerca con la acústica (Ridouane, 2011).
Análisis morfológico. En este nivel se estudia la composición de las palabras (Bolshakov & Gelbukh, 2004). Por ejemplo, la palabra "deshacer" está compuesta por el prefijo "des" y el verbo "hacer". Un humano puede descomponer las palabras en sus diferentes compuestos y a partir de ellos interpretar el significado de una palabra hasta entonces desconocida.
Análisis léxico. En este nivel, se análiza el significado de cada palabra de forma individual.
Análisis sintáctico. En este nivel se analiza la estructura de las palabras que forman una oración.
Análisis semántico. El análisis semántico determina el posible significado de una oración basado en la relación entre las palabras que la componen. En este nivel se realiza la desambiguación de las palabas que pueden adoptar difentes significados. Una palabra ambigua es "papa" que al omitir el uso de mayúsculas y minúsculas bien puede referirse a un tubérculo comestible o al lider cátólico.
Análisis del discurso. En este nivel se análiza el significado no una oración, si no de un texto completo, haciendo conecciones entre los significados de las oraciones individuales.

4. Aplicaciones del procesamiento de lenguaje natural

Cada aplicación del NLP puede utilizar uno o varios de los niveles mencionados. Dentro de las principales aplicaciones del NLP se pueden mencionar:

La recuperación de información

El mayor tesoro que tiene la humanidad es el conocimiento. Durante muchos años la actividad principal del hombre ha sido producir, guardar y transmitir conocimiento. La mayor parte de este conocimiento se almacena y transmite en forma de lenguaje humano. El acceso a esta información de manera eficiente y eficaz se ha vuelto más necesario a medida que la cantidad de conocimiento al que tenemos acceso es mayor. De allí que la aplicación del NLP más conocida e importante se la recuperación de información (Gelbukh & Sidorov, 2010). La recuperación de información tiene como objetivo encontrar recursos relevantes a partir de grandes colecciones para suplir una necesidad de información (Baeza-Yates & Ribeiro-Neto, 2011).

La extracción de información

La extracción de información tiene como objetivo reconocer, etiquetar y extraer datos estructurados a partir de texto no estructurado (Chang, et al., 2006; Freitag, 2000). La identificación de algunos elementos claves del texto como personas, lugares, organizaciones es muy importante para las aplicaciones de minería, visualización o en la respuesta de preguntas (Liddy, 2001).

Respuestas a preguntas

A pesar de que la información esté disponible para leer y encontrar respuesta, se necesitan sistemas que permitan al usuario realizar una pregunta en su propio lenguaje y recibir una respuesta rápida y apropiada (Hirschman & Gaizauskas, 2001). Los buscadores tradicionales dan una lista de los posibles documentos relevantes a su inquietud, en cambio los sistemas de respuestas a preguntas le ofrecen al usuario el texto de la respuesta o porciones de texto que contiene la respuesta (Liddy, 2001).

Resumen automático

El resumen automático del texto genera una versión condensada del documento que contiene unas pocas oraciones importantes (Liddy, 2001; Gambhir & Gupta, 2016), así un resumen automático debe consistir de la información más importante del documento

y al mismo tiempo, ocupar menos espacio que el documento original (Gambhir & Gupta, 2016). Esta tarea surge a mediados del siglo pasado (Vodolazova, et al., 2013) y desde entonces los investigadores buscan la forma de mejorar la calidad del resumen automático, apareciendo técnicas basadas en la estadística, en el análisis temático, en el estudio de los grafos, en el análsis del discurso y aprendizaje de máquina (Gambhir & Gupta, 2016).

Traducción automática

La traducción automática de texto en lenguaje natural es una aplicación de la lingúistica computacional que tiene como objetivo estudiar el uso de software para traducir un texto, en forma oral o escrita a otro lenguaje. Se han utilizado varios niveles en la traducción automática, desde la traducción "palabra por palabra", hasta análisis más complejos.

Para el estudio de la traducción automática se requiere tener grandes cantidades de texto paralelo, es decir, el mismo texto en lenguajes diferentes. A menudo, la fuente para estos textos paralelos son comunmente instituciones multinacionales como las Naciones Unidas o la Unión Europea, que producen los documentos para los diferentes miembros en lenguaje diferente (Koehn, 2005).

Para el estudio y evaluación de las diferentes aplicaciones del procesamiento de lenguaje natural se utilizan colecciones de prueba o corpus de documentos. El texto contenido en tales colecciones es generado por humanos y se utiliza para experimentación y comparación de diferentes técnicas. En este artículo se con analiza la creación de un corpus general basado en noticias publicadas en periódicos digitales de Honduras.

5. Corpus

En la actualidad, la disponibilidad de datos condiciona la investigación en procesamiento de lenguaje natural. Aunque en el país este tipo de investigación es escasa, se busca proveer a los docentes e investigadores interesados una colección de documentos escritos en Honduras. La construcción de un corpus de noticias publicadas en medios digitales de Honduras requiere de algunos retos importantes. En general, fue necesario:

- Analizar la composición de los sitios web de los diferentes medios digitales de Honduras. La principal carateística estudiada fue la forma en que se organizan digitalmente las direcciones URL (*uniform resource locator*). Las URL con un patrón claramente definido permiten la creación facil y eficiente de un *web crawler* personalizado.
- Programar de un *web crawler* personalizado. Un *web crawler* es un pieza de software que permite navegar de forma automática o semiautomática los sitios web de interés.
- Crear una copia local del texto contenido en las noticias de los sitios web seleccionados para la creación del corpus.

La colección de documentos fue procesada para eliminar el marcaje html (*hipertext markup language*) original del sitio web de origen. Además del contenido de la noticia se identificó el título, la fecha, el autor, la sección, la dirección URL y el nombre del periodíco digital del que se tomó. Finalmente, se unificó todas las noticias en un solo documento XML (*extensible markup languaje*), marcando las etiquetas con los datos mencionados. En la sección de resultados se describe la el formato y la colección resultante.

Resultados y discusión

El primer reto al construir un corpus es la selección de fuentes de datos para la colección. (Baker, et al., 2004). En este artículo se describe la colección UTD-MB-2016 que utiliza como fuente las noticias publicadas en el sitio web de La tribuna (La Tribuna, 2009),

Diario deportivo más (Diario Deportivo Más, 2010)y Diario Tiempo (Diario Tiempo Digital, 2015). Se determinó utilizar La tribuna y Diario Tiempo como fuente de noticias generales, con temáticas y secciones varias. En cambio, Diario Deportivo Más es un diario especializado en noticias deportivas. Esta composición favorece el estudio y aplicación de ciertas tareas del procesamiento de lenguaje natural.

Formato

La Figura 1 muestra el formato XML utilizado para almacenar la colección de noticias. Para cada noticia se define el título, la URL, el autor, la fecha, la fuente, la sección y el contenido. El título es el asignado por el autor de la noticia. En el campo URL se guarda la dirección donde puede encontrarse la noticia. El campo autor, guarda el nombre de el o los autores identificados en la noticia. La fecha se refiere a la fecha de publicación de la noticia. La fuente es el nombre del periódico del que fue extraída la noticia. La sección se extrae si está claramente definida en el texto de la noticia y por el último el contenido.

Figura 1. Formato XML de la colección.

```
<noticias>
    <noticia>
        <titulo></titulo>
        <url></url>
        <autor></autor>
        <fecha></fecha>
        <fuente></fuente>
        <seccion></seccion>
        <contenido></contenido>
    </noticia>
</noticias>
```

Fuente: Elaboración propia.

Distribución

La distribución de las noticias por fuente se muestra en la tabla 1. En total, se recuperaron 178125 noticias, de las cuales el 63% fueron tomadas de La tribuna. El 20.7% provienen del diario Deportivo Más y y el 16.3% de diario Tiempo. En promedio, por cada noticia se tienen 1714 bytes. En relación al tamaño de almacenamiento, la colección tiene 291.2 MB de contenido de noticias. Manteniendo la proporción relativa a la cantidad de noticias, la mayor parte del contenido, 178.9 MB, provienen de La Tribuna.

Tabla 1 – Distribución de las noticias por fuente

Fuente	*Noticias*	*Tamaño promedio*	*Tamaño total*
Diario deportivo Más	*36805*	*1703.48*	*59.8 MB*
La Tribuna	*112293*	*1670.83*	*178.9 MB*
Tiempo	*29027*	*1895.12*	*52.5 MB*
Total	*178125*	*1714.12*	*291.2 MB*

Fuente: Elaboración propia.

En relación a la producción diaria, el diario deportivo Más publica en promedio 31.67 noticias diarias. De nuevo, La Tribuna es quien más produce con 132.27 noticias diarias y el diario Tiempo produce en promedio 71.67 noticias por día. Estas estadísticas y otros referentes se pueden ver en la tabla 2.

Tabla 2 – Cantidad de noticias diarias para cada fuente

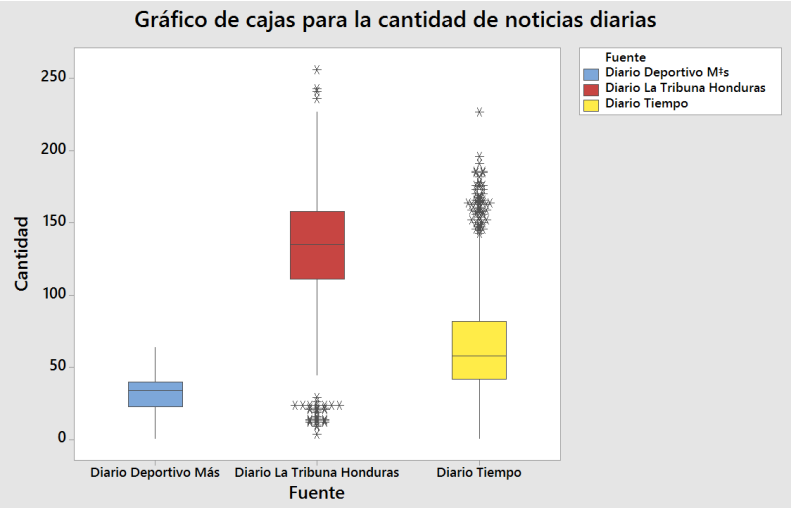
Fuente	*Media diaria*	*Desviación estándar*	*Minimo*	*Máximo*
Diario deportivo Más	*31.673*	*11.57*	*1*	*64*
La Tribuna	*132.27*	*37.41*	*4*	*256*

Tiempo	*71,67*	*44.91*	*1*	*227*

Fuente: Elaboración propia.

La proporción de noticias diarias se puede notar gráficamente en la **Figura 2**. Como se puede notar, igual que en la tabla 2, la mayor cantidad de noticias diarais fueron producidas por La Tribuna. Sin embargo, tanto La tribuna, como el Tiempo, poseen una gran cantidad de puntos estadísticamente muy desiguales, denotados en el gráfico por los asteriscos que aparecen fuera de la caja. Estos asteristicos son valores atípicos que se alejan de la media diaria. En cambio, el diario deportivo Más, parece ser más constante, denotado en el gráfico por la caja más pequeña, lo que indica que la cantidad de noticias está casi siempre alrededor de la media. También puede notarse que para este diario, no hay valores estadísiticamente atípicos.

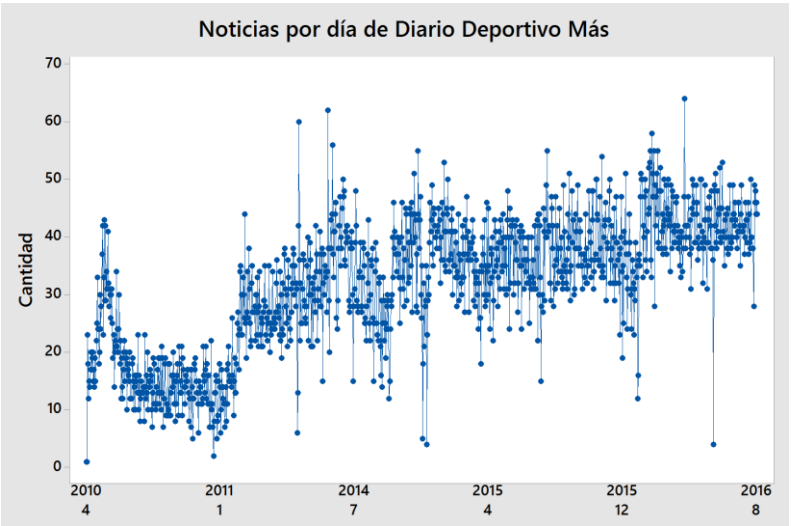
Figura 2. Gráfico de cajas para la cantidad de noticias diarias por cada fuente.



Fuente: Elaboración propia.

El Diario Deportivo Más tiene su primera publicación el 22 de abril de 2010. Apartir de entonces ha comenzado a aumentar su producción diaria, como se puede ver en la **Figura 3**. Un salto importante se puede no tar en el año 2011, como puede verse, no hay datos para el 2012 y 2013. Durante estos años el Diario Deportivo Más no realizó publicaciones en línea.

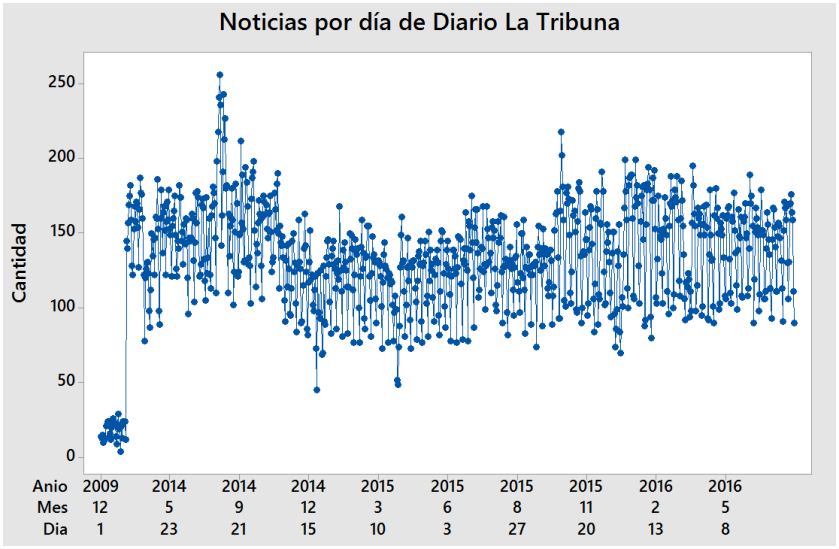
Figura 3. Serie de tiempo de la cantidad de noticias por día del Diario Deportivo Más.



Fuente: Elaboración propia.

Por su parte, La Tribuna ha realizado una mayor cantidad de noticias por día (ver **Figura 4**). Sin embargo, aunque comenzó su versión electrónica en los finales del año 2009, no se recuperaron noticias que hayan sido publicadas entre los años 2010 al 2013. Es hasta marzo de 2014 que retoman la versión en línea que se mantiene constante hasta la fecha en que se realizó esta investigación.

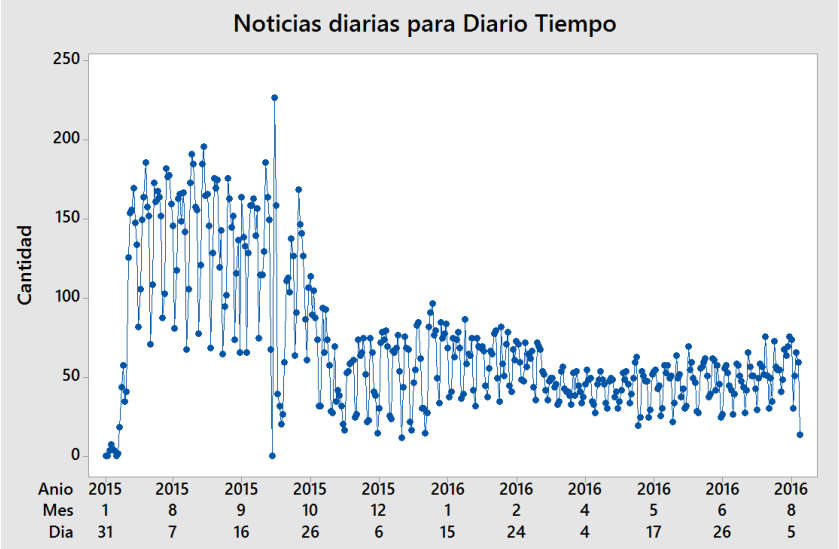
Figura 4. Noticias por día de Diario La Tribuna



Fuente: Elaboración propia.

Finalmente, Diario Tiempo es el que más recientemente tiene versión electrónica. Comenzó a publicar en su sitio web a partir del año 2015. Además, como puede verse en la **Figura 5**, para finales del 2015 decayó el volumen de las publicaciones diarias.

Figura 5. Noticias por día publicadas por Diario Tiempo



Fuente: Elaboración propia.

Conclusiones

En los últimos años, los periódicos hondureños han comenzado a tener una presencia fuerte en sus versiones electrónicas. Con el objetivo de fortalecer la investigación en el área de lingüística computacional y específicamente en el procesamiento de lenguaje natural, como resultado de esta investigación se propone el uso del corpus creado a partir

de noticias publicadas en las versiones electrónicas de tres diarios hondureños: La tribuna, Diario Deportivo Más y Diario Tiempo. La colección UTD-MB-2016 se almacena en formato XML para proveer una forma estandar de acceder a las noticias. Además se etiquetan los metadatos más importantes que podrán ser utilizados en varias de las tareas del procesamiento de lenguaje natural. La colección posee cerca de 300 MB de contenido, siendo La Tribuna quien aporta una mayor cantidad de noticias. Las noticias están fechadas a partir del 2009 y hasta agosto de 2016 que se creó la colección. Sin embargo, no hubieron publicaciones en algunos de los años intermedios. Los investigadores esperan que el trabajo realizado en la conformación de esta colección de noticias hondureñas, pueda motivar la investigación y enseñanza de un tema tan actual, como lo es el procesamiento computarizado del lenguaje humano.

Referencias

Chowdhury, G. G., 2003. Natural language processing. *Annual review of information science and technology,* 37(1), pp. 51-89.

Liddy, E. D., 2001. Natural Language Processing. En: *Encyclopedia of Library and Information Science.* 2nd Edition. ed. New York: Marcel Decker.

Ridouane, R., 2011. The phonetics and phonology of geminate consonants. *NINJAL international conference on phonetics and phonology (ICPP 2011).*

Gelbukh, A., 2010. Procesamiento de lenguaje natural y sus aplicaciones. *Korpus Sapiens. Sociedad Mexicana de inteligencia artificial,* Volumen 1, pp. 6-11.

Bolshakov, I. A. & Gelbukh, A., 2004. *Computational linguistics models, resources, applications.* 1ra. Edición ed. Mexico: Ciencia de la Computación.

Gelbukh, A. & Sidorov, G., 2010. *Procesamiento automático del español con enfoque en recursos léxicos grandes.* 2da edición ed. México: Instituto Politécnico Nacional.

Baeza-Yates, R. & Ribeiro-Neto, 2011. *Modern Information Retrieval.* 2da edición ed. New York: ACM press New York.

Chang, C.-H., Kayed, M., Girgis, M. R. & Shaalan, K. F., 2006. A survey of web information extraction systems. *IEEE transactions on knowledge and data engineering,* 18(10), pp. 1411-1428.

Freitag, D., 2000. Machine learning for information extraction in informal domains. *Machine learning,* 29(2-3), pp. 169-202.

Hirschman, L. & Gaizauskas, R., 2001. Natural language question answering: the view from here. *Natural Language Engineering,* 7(4), pp. 275-300.

Koehn, P., 2005. *Europarl: A Parallel Corpus for Statistical Machine Translation.* s.l., s.n., pp. 79-86.

Vodolazova, T., Lloret, E., Muñoz, R. & Palomar, M., 2013. *Extractive Text Summarization: Can We Use the Same Techniques for Any Text?.* s.l., Springer, pp. 164-175.

Gambhir, M. & Gupta, V., 2016. Recent automatic text summarization techniques: a survey. *Artificial Intelligence Review,* pp. 1-66.

Baker, P. y otros, 2004. Corpus linguistics and South Asian languages: Corpus creation and tool development. *Literary and Linguistic Computing,* 19(4), pp. 509-524.

La Tribuna, 2009. *La Tribuna.* [En línea]
Available at: http://www.latribuna.hn
[Último acceso: Agosto 2016].

Diario Deportivo Más, 2010. *Diario Deportivo Más.* [En línea]
Available at: http://www.diariomas.hn
[Último acceso: Agosto 2016].

Diario Tiempo Digital, 2015. *Diario Tiempo Digital.* [En línea]
Available at: http://tiempo.hn
[Último acceso: Agosto 2016].

Autorización y Renuncia

Los (a) autores facultan a CEAT para publicar el escrito en los procedimientos de la conferencia. CEAT o los editores no son responsables por el contenido y las implicaciones de lo que esta expresado en el escrito.