

Glosario de términos.

Cooperativa
CubaSWL

ToNgueLP, editor de corpus de textos para tareas NLP.

Producto: ToNgueLP Corpus Tools

Versión: 1.0

Autor: Ing. Abel Meneses Abad

Reglas de Confidencialidad

Clasificación: Software de licencia dual libre con fines de investigación, y comercial para su aplicación en ámbitos con fines de lucro, factible de exportar por las partes desarrolladoras, de carácter nacional.

Este documento contiene información propietaria de **Abel Meneses Abad**, y es emitido confidencialmente para orientar al equipo de desarrollo de ToNgueLP.

El que recibe el documento asume la custodia y control, comprometiéndose a no reproducir, divulgar, difundir o de cualquier manera hacer de conocimientos público su contenido, excepto para cumplir el propósito para el cual se ha generado.

Estas reglas son aplicables a las 8 páginas de este documento.

Índice de contenidos

Glosario de términos.....	1
Referidos a la Lingüística.....	2
Disciplinas Lingüísticas.....	2
Relaciones Semánticas.....	2
Paráfrasis.....	2
Otros.....	3
Referidos al Procesamiento del Lenguaje Natural.....	4
Referidos al Procesamiento de Textos.....	6
Referidos a la Detección de Plagio.....	7
Bibliografía.....	8

Referidos a la Lingüística

Disciplinas Lingüísticas

Esta clasificación mayoritariamente viene de (Jurafsky & Martin, 2009)

- **Fonética & Fonología** (Phonetics & Phonology):
- **Morfología lingüística:** estudia la forma de las palabras, estudia las palabras y su estructura interna.
- **Lexicología:**
- **Semántica:**
- **Sintaxis:** describe como las palabras se combinan para formar sintagmas, oraciones y frases.
- Pragmática
- Discurso

Relaciones Semánticas

Hipónimo: Los hipónimos son palabras o términos específicos de una clase o conjunto de significado más amplio o general (hiperónimo). Término del ámbito semántico, algunos de sus ejemplos son 'coche' o 'moto' con respecto a 'vehículo', 'rosa' o 'margarita' en relación con 'flor', 'cerveza' o 'vino' y 'bebida', etc.

Tropónimos: son hipónimos verbales que denotan una determinada manera o modo de llevar a cabo la acción del verbo. El término proviene de las palabras griegas tropos, que significa 'modo, manera', y ónoma, 'nombre'. Es el caso, por ejemplo, de susurrar (hipónimo de hablar, pero en voz baja, a modo de murmullo), jugar (pero sin propósito determinado), deambular (caminar, andar, pero sin dirección determinada), etc.

Transcreation (Transcreación): es la re-escritura completa de un texto en un idioma extranjero teniendo en cuenta las especificidades culturales de dicha lengua. El término transcreación es un componente esencial en la internacionalización. Se aplica generalmente en marketing y publicidad donde el contenido de la comunicación debe producir el mismo impacto en el público del mercado al que nos dirigimos que el contenido original en el idioma original.

Paráfrasis

- Same Polarity
- Order

- Synthetic/Analytic
- Semantic
- Spelling
- Addition/Deletion

Otros

Lexicón: Conjunto abstracto no ordenado de entradas léxicas que se definen de acuerdo a sus rasgos fónicos y gramaticales. / Conjunto de las palabras y lexemas de una lengua y libro en que se contienen. (Moreno, 2000)

Flexión (en lingüística): es la alteración que experimentan las palabra mediante morfemas constituyentes según el significado gramatical o categórico para expresar sus distintas funciones dentro de la oración y sus relaciones de dependencia o de concordancia con otras palabras o elementos oracionales. (Wikipedia en español) Ejemplo los sufijos flexivos al cambiar la categoría(o función léxica) de una palabra expresan esta propiedad; también lo hacen los morfemas cuando cambian número o persona.

verbos modales:

diátesis:

elipsis (en lingüística):

morfemas: letras que pospuestas al lexema indican los accidentes del vocablo: género, número, tiempo y persona.

afijo: el conjunto que reúne los sufijos, prefijos, interfijos, infijos y circunfijos.

Acento tónico: cuando una sílaba recibe la acentuación se dice que es tónica.

Referidos al Procesamiento del Lenguaje Natural

Natural Language Processing (NLP): Procesamiento del Lenguaje Natural, área de la computación, altamente relacionada con la lingüística, que engloba todos los problemas de entendimiento (understanding) o generación (generation) del lenguaje humano.

Fuzzy string matching: ...(extraído de Taming Text)

Lematizar (lemmatiz): proceso mediante el cual se extrae el “**lexema**” de la palabra. Generalmente es necesario utilizar una base de datos que contenga información de los lexemas o lemas (como también se le suele llamar a los lexemas), estas BD son generalmente semánticas. Trabaja la morfología de las palabras.

Stemming: proceso mediante el cual se eliminan de la palabra los “**morfemas**”, utilizando reglas predefinidas que se corresponden con las terminaciones más comunes de las palabras en un idioma. Trabaja la morfología de las palabras.

Compresión lingüística: procesos del tratamiento lingüístico en el que se reducen los términos estudiados. Ejemplo: la lematización y el stemming son tipos de compresión lingüística.

Part of Speech (POS): área de NLP que estudia el como encontrar las funciones léxicas de las palabras o las partes de la oración, dentro de las dos formas en que se presenta el lenguaje natural (idiomas humanos), los textos y ficheros de audio.

Entity identification (or entity recognition): es un problema de etiquetado dentro de NLP que consiste en identificar las entidades existentes en un texto, como por ejemplo: lugares, nombres, etc.

Lexicón computacional: subconjunto del conjunto total de entradas que constituye un lexicón para una lengua natural, normalmente un dominio específico.

n-gramas (n-grams): subconjunto de n fragmentos, que pueden ser caracteres o palabras. Ej: La computadora estaba encendida. 3-gramas: “La computadora estaba”, y “computadora estaba encendida”, tiene dos 3-gramas, y un solo 4-grama.

Chunking: cortar en pedazos largos. En lingüística significa tomar segmentos largos de varias frases para analizarlas. O sea dividir en pasajes, que pueden superponerse incluso, sin que estos tengan que coincidir obligatoriamente con los párrafos del texto. Detalles: se menciona en la literatura el término “chunking strategy” ejemplo de ellas pueden ser: (i) 25-line chunking, (ii) 4-sentence chunking or (iii) paragraph chunking.

textual entailment (implicación textual): en la lógica el texto 1 implica el texto 2, esto podría ser el texto 1 es la pregunta que responde el texto 2; u otro tipo de relaciones.

Token: A token is the technical name for a sequence of characters — such as hairy, his, or :) — that we want to treat as a group. (Bird 2009, NLP with Py)

Semantic Role Labeling: identifying how a noun phrase relates to the verb (as agent, patient, instrument, and so on). Para cada clausula , determinar el rol semántico jugado por cada frase sustantiva que es un argumento del verbo. También se refiere a este tema como “case role analysis,” “thematic analysis,” and “shallow semantic parsing” Roles semánticos más comunes:

- Agent: Actor of an action
- Patient: Entity affected by the action
- Instrument: Tool used in performing action.
- Beneficiary: Entity for whom action is performed
- Source: Origin of the affected entity
- Destination: Destination of the affected entity

Parse Trees: árboles de análisis gramatical. Resultado de la descomposición de una oración en sus partes componentes, tales como: sujeto y predicado, y luego estos en componentes más específicos y así sucesivamente.

Referidos al Procesamiento de Textos

Corrección de errores (normalización de textos): procesos del tratamiento lingüístico que elimina errores a partir de la comparación con supuestos correctos de la lengua. Ejemplo: Camaguey lleva diéresis en la u(ü), su forma correcta sería: Camagüey.

Referidos a la Detección Automática de Plagio

Automatic Plagiarism Detection (Detección Automática de Plagio): disciplina de las ciencias de la computación que analiza, diseña, implementa y prueba algoritmos que resuelven los problemas de cada una de las fases de la detección de texto re-usado sin la cita adecuada de sus autores originales. Mayoritariamente se dedica a la detección del plagio en textos, existen pocos ejemplos de la utilización de estas técnicas en otras propiedades intelectuales digitales como la pintura digital, medias, etc. Estas técnicas pertenecen también a la ciencia forense de textos digitales. Suele dividirse en tres etapas: Source Retrieval, Text Alignment y Knowledge Based Post-Processing.

Metodos para Re-utilizar un Texto (Methods for Text Re-Use)(Barrón-Cedeño, 2012), pág13

-

Sparse matching: coincidencia esparcida, se aplica cuando dos documentos son similares en uno o más fragmentos aislados.

paraphrase plagiarism (plagio parafrástico): consiste en tomar los escritos de otra persona, añadiendo o quitando palabras o letras, agregando errores deliberados de ortografía y gramática, insertando palabras con significados similares o reordenando oraciones y frases.”(Kakkonen y Mozgovoy, 2010)

verbatim plagiarism: plagio literal o copy-paste.

high context similarity: similaridad en textos largos.

short context similarity: similaridad en textos cortos, como resúmenes, noticias, etc.

Umbrales(idea del 01/10/2013): La heurística pudiera considerar lo siguiente, 4 umbrales. El primero es la "similitud en la colección" (tf-idf, usado por Kong, hace esto)[valorar luego como insertar en este umbral el tipo de índice basado en citas(ver url de wikipedia que lo explica, o buscar más, escribir a Gi)]. 2do umbral "similitud entre dos documentos"(high threshold(ediciones por ejemplo) o lowthreshold(un fragmento o varios cuya \sum representa un pequeño % del documento total)). 3er umbral "posición" o sea si dos fragmentos son iguales su primera oración similar no dista de la última oración similar en más del 40% del total del fragmento. 4to umbral, "número de oraciones similares", tal y como refleja la literatura este umbral se puede hacer ≥ 3 , pues dos oraciones o una no refleja contenido suficiente para establecer una similitud considerable a texto reusado.

high similarity threshold: un gran % del documento similar a otro.

low similarity threshold: un pequeño % del documento similar a otro. Generalmente son similares en fragmentos, párrafos o textos cortos. No se refiere en ningún momento a similaridad en oraciones.

Semantic density:

Actually, AI / NLP is not even scientific: A science integrates all its disciplines. However, the disciplines of AI are incompatible. Thus, AI is not fundamental, and therefore not scientific. Example: Ontology (automated reasoning) and NLP are incompatible: Ontology is based on formal language and NLP is based on natural language. And therefore, scientists fail to integrate both to: reasoning in natural language.

Why is crucial information discarded during the NLP process? Crucial information makes the difference between the current quick-and-dirty approach to knowledge technology and my fundamental approach. Example: Scientists fail to implement automatic generation of useful questions - because of this crucial information that is lost - while generating useful question is extremely simple: My system utilizes Natural Laws of Intelligence in grammar; A Natural Law of Intelligence in grammar: Conjunction “or” defines a choice; Given “A person is a man or a woman.” and “Chris is a person.”; Substitution of both sentences: “Chris is a man or a woman.”; Conversion to a question: “Is Chris a man or a woman?”. (03/11/2013 a través de LinkedIn)

Bibliografía
