Cooperativa CubaSWL

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

Índice de contenidos

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), juguetear (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 reune los sufijos, prefijos, interfijos, infijos y circunfijos.

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

Engaño verbal (verbal deception):

Collocations: secuencia de palabras que aparecen juntas de forma frecuente, estableciéndose como nuevos códigos de la lengua. Ej. “caballero negro”, “vino blanco”, “Estados Unidos de Norteamérica”, etc.

Sigla: es el proceso de creación de palabras a partir de cada grafema inicial de los términos principales de una expresión compleja. Solo se separan por punto cuanto están en un texto en mayúsculas. Ej. **ONU**, **O**rganización de las **N**aciones **U**nidas. Ej. BOLETÍN DE LA O.E.A.

Acrónimo: sigla que se lee como una palabra o un vocablo formado al unir parte de varias palabras. Ej. **Radar**, del inglés ***Ra****dio* ***D****etection* ***a****nd* ***R****anging.*

Abreviatura*:* representación escrita de una o varias palabras mediante una o varias de sus letras, a fin de que la palabra o las palabras en cuestión resulten más cortas en el texto. Ej. antes de Cristo: **a. C.**, avenida: **av.**, páginas: **págs.**

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 (undestanding) o generación (generation) del lenguaje humano.

Tareas de NLP

- **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.
- **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. (Wikipedia)[1]
- **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.
- **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 compunes:
  - 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

Técnicas de Pre-procesamiento de Textos

- **Chunking**: cortar en pedazos largos. En linguística 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.
- **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.

1 http://en.wikipedia.org/wiki/Textual_entailment

- **Lematización** (lemmatization): 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.

Métricas de cadena

(String Metrics) Son las métricas que miden similaridad o no-similaridad entre cadenas de texto.(Wikipedia)[2]

- Levenstein Distance

Otros

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.

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.

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)

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.

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

word lattice: es un gráfico directo que representa eficientemente mucha más información sobre posibles secuencias de palabras.(Jurafsky2009, p. 372)

2 http://en.wikipedia.org/wiki/String_metric

Referidos al Procesamiento de Textos

Corrección Ortográfica (spell checking): 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.

Normalización de Textos (text normalization): es el subproceso que implica mezclar diferentes formas de escritura en una sola apropiada y aceptable; por ejemplo un documento puede contener los símbolos “Señor”, “señor”, “Sr.”,”Sr” todos ellos deben ser normalizados a una única forma. “...involves merging different written forms of a token into a canonical normalized form; for example, a document may contain the equivalent tokens “Mr.”, “Mr”, “mister”, and “Mister” that would all be normalized to a single form. ” (Indurkhya 2008, p. 10)

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. para Suele dividirse en tres etapas: Source Retrieval, Text Alignment y Knowledge Based Post-Processing.

Métodos para Reutilizar Textos

(Methods for Text Re-Use)(Barrón-Cedeño, 2012, pág-13)

-

Tipos de Plagio

(Plagiarism Tipologies)(Barrón-Cedeño, 2012, pág-17; Imran, 2010)

-

Métodos de Plagio

Como se implementan los tipos de plagio.

- **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.
- Instrinsic Plagiarism
- Extrinsic Plagiarism
- Mono-lingual Plagiarism
- Cross-lingual Plagiarism
- Idea Plagiarism

Métricas

Para evaluar el proceso de detección automática de plagio.(Oberreuter & Velásquez, 2012, p.8), (Grozea & Popescu, 2012, p.7), (Potthast, Stein, Barrón-Cedeño, & Rosso, 2010),

- Precision
- Recall
- Granularity
- Plagdet
- F-measure

Otros

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

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 treshold(ediciones por ejemplo) o lowtreshold(un fragmento o varios cuya ∑ 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 treshold: un gran % del documento similar a otro.

low similarity treshold: 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

Barrón-Cedeño, A. (2012). *On the Mono- and Cross-Language Detection of Text Re-Use and Plagiarism*. Universidad Politécnica de Valencia.

Grozea, C., & Popescu, M. (2012). Encoplot - Tuned for High Recall (also proposing a new plagiarism detection score). *PAN'2012* (p. 12). Berlín, Germany.

Imran, N. (2010). Electronic media, creativity and plagiarism. *ACM SIGCAS Computers and Society*, *40*(4), 25–44. doi:10.1145/1929609.1929613

Jurafsky, D., & Martin, J. H. (2009). *Speech and Language Processing: An introduction to natural language processing* (Second Edi., p. 1044). Prentice Hall.

Moreno, A. (2000). DISEÑO E IMPLEMENTACIÓN DE UN LEXICÓN COMPUTACIONAL PARA LEXICOGRAFÍA Y TRADUCCIÓN AUTOMÁTICA. *Estudios de Lingüística del Español*, *9*. Retrieved from http://elies.rediris.es/elies9/index.htm

Oberreuter, G. G., & Velásquez, J. D. (2012). *Exploring Text Features for Automatic Detection of Plagiarism* (p. 14). Santiago de Chile, Chile.

Potthast, M., Stein, B., Barrón-Cedeño, A., & Rosso, P. (2010). An Evaluation Framework for Plagiarism Detection. *Coling 2010, 23rd International Conference on Computational Linguistics* (pp. 997–1005). Beijing, China: ACM Press.

Potthast, M., Stein, B., & Rosso, P. (2010). An Evaluation Framework for Plagiarism Detection.