Evaluación del Módulo 3

Laura Rodríguez Navas
rodrigueznavas@posgrado.uimp.es

Febrero 2020

1. **Describa los distintos tipos de recursos lingüísticos que puede usar un sistema de PLN, así como indique al menos dos ejemplos de cada tipo de recurso. Se valoraran positivamente que los ejemplos se refieran a recursos para el procesamiento de Español (2 puntos).**

   Los tipos de recursos lingüísticos que puede usar un sistema de PLN son: Lexicones y Diccionarios, Ontologías y Corpus.

   Los **Lexicones** son series ordenadas de palabras de una lengua, una persona, una región, una materia o una época determinada. Y podemos clasificar-los como generales o especializados.

   Los lexicones generales son repositorios de palabras. Pueden ser tan simples como una lista de palabras o pueden ser tan complejos como una base de datos terminológica. También incorporan información o conocimiento sobre las palabras; como información fonológica, morfológica, sintáctica, semántica y pragmática.

   Hay de muchos tipos de lexicones especializados, dentro de distintas categorías, como, por ejemplo:

   - Diccionarios de locuciones.
   - Bases de datos léxicas.
   - Gazetteers.
   - Bases de datos terminológicas.
   - Listas de nombres propios.
   - Listas de siglas o jergas.
   - Detectores de fechas, números, fórmulas, etc.

   En realidad, los lexicones especializados están orientados a ciertas aplicaciones o a ciertos dominios concretos. Por ejemplo, en las aplicaciones de análisis de sentimientos aparecen listas de términos polares (positivos y negativos); y en las aplicaciones de recuperación de información geográfica se pueden tener listas de topónimos y localizaciones. Por otra parte, en los dominios de tipo biomédico aparecen listas de términos; y en los dominios de tipo turístico se pueden tener listas de monumentos, ciudades, etc.; o también listas de topónimos, como en el caso de las aplicaciones.

   Los **Diccionarios** son listas de palabras, donde cada palabra contiene su definición y a veces su etimología. Existen diferentes tipos de diccionarios dependiendo de si son normativos, de uso, de

aprendizaje, etc. Y según el número de lenguas utilizadas puede ser diccionarios monolingües, bilingües o multilingües.

Dentro de este tipo también se encuentran las enciclopedias y los tesauros que son recursos lingüísticos parecidos a los diccionarios, pero son un poco más avanzados. Por ejemplo, un tesauro es una lista de palabras o términos controlados, empleados para representar conceptos.

Una **Ontología** es la especificación formal y explícita de una conceptualización compartida.

- Conceptualización porqué es una forma de entender o describir un dominio.
- Compartida porqué se entiende consensuada por un grupo y por varias partes.
- Explícita porqué está descrita en un lenguaje.
- Formal porqué es comprensible por las máquinas.

Los tipos de ontologías conocidas son:

- Las genéricas, que describen conceptos genéricos (espacio, tiempo, objeto, ...).
- Las específicas para un dominio, que expresan conceptos de dominios particulares (medicina, bioquímica, turismo, ...).
- Las específicas para una tarea, que expresan conceptos sobre la resolución de problemas (diagnósticos, ventas, ...).
- Las que están diseñadas específicamente para una aplicación, que describen conceptos que dependen tanto de un dominio específico como de una tarea específica (procesos de producción, cardiología, ...).

En la actualidad, el uso de las ontologías y su integración en los sistemas de PLN se está potenciando muchísimo precisamente porqué están apareciendo nuevas tareas que requieren de su potencial para trabajar con información que puede ser ingerida o de sentido común y con ello poder generar y desarrollar nuevos sistemas mucho más eficientes. Alguna de esas nuevas tareas son la Web semántica, la implicación textual, la minería de datos y el Big Data.

Este uso de las ontologías es para:

- Definir e unificar un espacio global de la información para compartir conocimiento, facilitar el razonamiento automático, etc.
- Estandarizar tipos de datos.
- Simplificar el acceso a los datos.
- Gestionar los datos definiendo los conceptos y sus relaciones básicas para la comprensión de un área.

Un **Corpus** es una colección de textos representativos de una lengua, de un dialecto o un subconjunto de una lengua, que se utiliza para el análisis lingüístico. Si además incluye información lingüística adicional, representa una herramienta y un recurso fundamental con un valor añadido para integrarse en tareas del PLN. De hecho, son uno de los recursos lingüísticos más utilizados en el ámbito del PLN con la aparición de la Web y Internet; y el uso de estos se ha fomentado porqué proporcionan la obtención de recursos de manera fácil y accesible.

Los Corpus se pueden clasificar,

- según el material que incorporan como textuales o orales.

- según el propósito como generales o específicos.

  Los corpus con fines generales, tienen como objetivo principal, constituir una fuente de información textual de una lengua para fines y aplicaciones diversas.

  En cambio, los corpus con fines específicos, se han creado como respuesta a un propósito particular, como el estudio de aspectos concretos de la gramática o del léxico de la lengua para un dominio concreto, para una aplicación concreta, o para una tarea concreta.

- según el número de lenguas utilizadas como monolingües o multilingües. Dentro de los corpus multilingües hay que distinguir entre los corpus paralelos y los corpus comparables.

  Los corpus paralelos son un conjunto de textos donde cada uno de éstos es la traducción exacta y fiable del idioma original a otras lenguas.

  Los corpus comparables son un conjunto de textos en varios idiomas que contienen información sobre un tema común pero que no requieren una traducción exacta del idioma original a otras lenguas.

- según la información lingüística que incorporan como anotados (etiquetados) o no anotados (no-etiquetados). Los corpus no anotados pueden ser recolecciones directas desde Internet. Y los corpus anotados son fundamentales ya que proporcionan una información valiosísima adicional en forma de marcas o anotaciones incluidas en cada secuencia de caracteres de los textos para trabajar con los sistemas del PLN.

Ejemplos

Lexicones

Dos ejemplos de lexicones en español que merecen especial atención son WordNet y Acquilex, que son considerados bases de datos léxicas multilingüe.

WordNet es un sistema electrónico de referencia léxica multilingüe, desarrollado en forma de base de datos léxica. Su diseño está en consonancia con teorías psicolingüísticas relativas a la organización de la información léxica en la mente del hablante. Además constituye el intento de reflejar un modelo de memoria léxica, basado en redes semánticas, en un modelo lexicográfico de organización léxica.

Los objetivos principales de WordNet son:

- La validación de las teorías psicolingüísticas sobre organización léxica anteriormente mencionadas.
- Su previsible utilización en diversas aplicaciones que requieran acceso a información léxica.

Concretamente, el WordNet en español que sigue el marco de EuroWordNet, se estructuran de la misma manera que el WordNet americano para el inglés (Princeton WordNet) en términos de conjuntos de sinónimos de palabras con relaciones semánticas básicas entre ellos. Los sustantivos, verbos y adjetivos en español se organizan en conjuntos de sinónimos, cada uno de los cuales representa un concepto léxico subyacente. Y diferentes relaciones son las encargadas de vincular los conjuntos de sinónimos.

La diferencia básica entre éste y otros lexicones es que pertenece a un único proyecto a gran escala en el que se ha tenido como idea fundamental la organización léxica en campos semánticos.

Por otra parte, **Acquilex** se basa en la utilización de diccionarios de soporte magnético (M.R.Ds, Machine Readable Dictionaries) para la construcción de componentes léxicos. Los diccionarios automatizados constituyen una fuente de adquisición de información léxica y conceptual que, potencialmente, permite abordar algunos aspectos especialmente costos de la construcción de una base de datos léxica multilingüe a partir de estos.

La extracción de la información de los MRDs está formada por:

- Análisis sintáctico de las definiciones.
- Extracción de información semántica de las definiciones analizadas.
- Desambiguación del género contenido en las definiciones y la construcción de taxonomías.
- Filtrado de la información contenida en la parte correspondiente a la diferencia específica de las definiciones.
- Conversión de los resultados de los procedimientos de extracción en un sistema de representación formal.

Acquilex se encuentra en un área poco explorada, debido a la dificultad que supone el tratamiento complejo de grandes volúmenes de información en el proceso de conversión de los MRDs a una base de datos léxica.

Diccionarios, Enciclopedias, Tesauros

En esta sección se destacan tres ejemplos. Un ejemplo de diccionario (el diccionario de la lengua española), un ejemplo de enciclopedia (la Wikipedia en español) y un ejemplo de tesauro (el ISOC de Economía).

- El diccionario de la lengua española es un diccionario normativo del idioma español editado y elaborado por la Real Academia Española (RAE). El diccionario incluye palabras de uso común extendido, al menos en un ámbito representativo de entre aquellos en los que se habla el español o castellano y además incluye numerosos arcaísmos y vocablos hoy en desuso, para entender la literatura castellana antigua. Actualmente, es la obra lexicográfica académica española por excelencia y se ha convertido en el diccionario de referencia y consulta del español.
- La Wikipedia es una enciclopedia `libre` [1] políglota y editada de manera colaborativa, que permite la recopilación, el almacenamiento y la transmisión de la información de forma estructurada. Una de sus ediciones es en español.

  La Wikipedia está sustituyendo a todas las enciclopedias españolas conocidas, como por ejemplo a la Enciclopedia Libre Universal en Español o la enciclopedia Encarta de Microsoft, aunque a veces su fiabilidad es cuestionada ya que puede ser editada por todo el mundo.

  Un dato importante en términos de PLN es que es una enciclopedia que se incorpora muy bien con los sistemas de PLN, con resultados realmente sorprendentes.

[1] no posee restricciones legales significativas en relación con el derecho de uso, la redistribución y la creación de versiones modificadas o derivadas por parte de terceros.

- El tesauro ISOC de Economía tiene como objetivo, facilitar un análisis homogéneo de los documentos que se incorporen a las bases de datos terminológicas y permitir al usuario la recuperación de estos de forma precisa y exhaustiva. El rasgo determinante de este tesauro es su base empírica y su utilidad contrastada durante décadas en distintos sistemas de información económica especializada. Por lo demás, su estructura interna responde al modelo tradicional de los tesauros que articula los términos en torno a una red de relaciones jerárquicas, asociativas y de equivalencia, eliminando así las posibles ambigüedades y polisemias, y facilitando la percepción de la afinidad y/o diferenciación semántica entre ellos.

  En concreto, este tesauro contiene 6.792 términos (5.464 descriptores y 1.328 no descriptores) distribuidos en 13 áreas temáticas y consta de un índice alfabético, otro jerárquico y un tercero permutado. Se completa con dos anexos: uno de identificadores, que recoge nombres de personas, instituciones, partidos políticos, etc.; y otro con la denominación de las monedas nacionales.

Ontologías

De ejemplos de ontologías hay muchísimos.

Curiosamente, el lexicón **WordNet**, comentado anteriormente, también puede ser un ejemplo de ontología ya que muchos de sus autores lo consideran como una auténtica ontología. De hecho porqué las ontologías generan cierta controversia en el campo de la Inteligencia Artificial (IA), ya que no existe claridad entre las ontologías y los lexicones.

Otro ejemplo parecido a WordNet, es BabelNet . **BabelNet** es una red semántica multilingüe creada a partir de la integración de WordNet y Wikipedia, y también una ontología lexicalizada. BabelNet ha sido creado automáticamente integrando la Wikipedia en la base de datos léxica WordNet. La integración se realiza a través de un mapeo automático, al tiempo que se llenan los huecos de idiomas con pocos recursos con la ayuda de traductores automáticos. El resultado es un diccionario enciclopédico que provee conceptos y entidades lexicalizadas en muchos idiomas, incluido el español, y conectadas a través de relaciones semánticas.

Su uso mayoritario es en aplicaciones de sistemas multilingües de PLN. Y el conocimiento lexicalizado que se encuentra disponible en BabelNet también se ha utilizado para obtener los mejores resultados conocidos hasta el momento en similitud semántica y desambiguación multilingüe.

Familiarmente a WordNet, BabelNet agrupa palabras de diferentes idiomas en conjuntos de sinónimos. Para cada conjunto de sinónimos, BabelNet provee pequeñas definiciones en muchos idiomas obtenidos tanto de WordNet como de Wikipedia.

Corpus

Dos ejemplos de corpus que incorporan el español son el corpus oral Albayzin y el corpus anotado 3LB.

El corpus oral español **Albayzin** consta de 3 subcorpus de señales de 16 kHz y 16 bits, grabadas por 304 hablantes de castellano.

Los 3 subcorpus son:

- un corpus fonético con 6.800 enunciados de oraciones fonéticamente equilibradas, incluidos 1000 enunciados con segmentación fonética.
- un corpus geográfico con 6.800 enunciados de oraciones extraídas de una base de datos geográfica española.
- un corpus "lombardo" con 2.000 declaraciones de diferentes corpus.

Albayzin fue producido en 1998 después de ser diseñado entre los años 1991 y 1993 por un consorcio de 6 grupos de investigación españoles liderados por el grupo de Procesamiento del Habla de la UPC.

Por otro lado, el corpus anotado **3LB** es un proyecto para la construcción de una base de datos de árboles sintáctico-semánticos que tiene como objetivo la construcción de tres corpus anotados sintácticamente para el español, el catalán y el euskera.

A pesar de que la construcción de tres corpus anotados sintácticamente es una tarea costosa, es una labor imprescindible para el desarrollo de aplicaciones reales en el área del PLN y como tal para el desarrollo de la sociedad de la información. Resulta imprescindible la obtención de gramáticas computacionales a partir de corpus que son un primer paso hacia procesos posteriores que requieren más elaboración. Entre estos procesos se halla la delimitación de las entidades discursivas, lo que, junto con la identificación de los elementos anafóricos y correferentes mejora sustancialmente la calidad de todos los sistemas de Traducción Automática (TA), Extracción de Información (EI), Recuperación de Información (RI), Resumen Automático (RA) y sistemas de Pregunta-Respuesta (PR).

Otras tareas lingüísticas que pueden abordarse si se dispone de tres corpus anotados sintácticamente son el aprendizaje de restricciones de selección o el de los patrones de subcategorización de los verbos.

2. Explique las aplicaciones de PLN que pueden estar interesadas en el uso de un Gazetter (2 puntos).

3. Describa las diferencias entre un reconocedor de entidades y un clasificador de entidades (2 puntos).

Un reconocedor de entidades solo detecta las palabras, dentro de un documento, que se comportan como entidades sin indicar de que tipo son. En cambio, un clasificador de entidades además de detectar las palabras que se comportan como entidades, clasifica las entidades reconocidas según su tipo.

Un clasificador de entidades también incorpora otra subtarea que un reconocedor de entidades no utiliza, que consiste en el proceso de la desambiguación de las entidades reconocidas.

4. Indique al menos 5 ejemplos de entidades ambiguas (2 puntos).

- La Universidad de Barcelona, ya que puede ser de tipo organización (ORG) o localización (LOC).
- La ciudad de París, ya que puede ser de tipo localización (LOC) como capital de Francia, o de tipo personas (PER) como Paris Hilton.
- Sun, ya que puede ser de tipo organización (ORG) como la empresa informática, o de tipo personas (PER) como el personaje de ficción de la serie Perdidos.

- La gran manzana, ya que puede ser de tipo localización (LOC) como la Ciudad de Nueva York, o de tipo miscelánea (MISC) como una manzana de gran tamaño.
- Floyd, ya que puede ser de tipo localización (LOC) como la población de Iowa, o de tipo personas (PER) como el grupo de música Pink Floyd.

5. Razone si el uso de un reconocedor de entidades sería beneficioso para el Análisis de Opiniones. El Análisis de Opiniones es la tarea que se encarga de la clasificación automática de opiniones (2 puntos).