Procesamiento del Lenguaje Natural

Guía Introductoria

Mario Alberich

Junio 2007

Tabla de contenidos

Comentarios previos

La intención principal que ha llevado a la redacción de esta guía introductoria ha sido exponer los aspectos clave del PLN incluyendo la terminología científica que se utiliza, en la medida de lo necesario. La lectura de este documento debe entenderse como una tarea introductoria y como paso previo a poder recopilar más información al respecto.

Este documento tiene una mera intención introductoria sobre el procesamiento del lenguaje natural, y por ello es poco exhaustivo, académicamente informal e incluso poco estricto en el uso de la terminología.

La intención principal del documento es pues, exponer de una forma lo más llana posible los aspectos clave del PLN, esperando que sea de utilidad.

En concreto, esta versión del documento se encuentra disponible en el apartado:

http://www.sopadebits.com/descargas#pln

Última revisión: 30/6/2007

Introducción

Origen, enfoque y aplicaciones

Disciplinas fuente del PLN

El procesamiento del Lenguaje Natural es una disciplina que relaciona directamente la informática y con la lingüística.

Su objetivo, es poder conseguir que el lenguaje humano (por contraposición a lenguajes de programación utilizados en máquinas) pueda utilizarse como entrada (*input*) en un proceso automatizado.

A partir de esta entrada de información se realiza:

- El procesamiento de la entrada, basándose en la gramática y los recursos lingüísticos, utilizando analizadores y estableciendo los límites y criterios de calidad.
- El procesamiento de la salida y la propia salida de datos, que varía según el ámbito de aplicación del PLN.

> El término **Procesamiento del Lenguaje Natural** (PLN) se aplica en el ámbito informático. Dentro de la lingüística, esta disciplina se denomina **Lingüística Computacional**.

Estas dos denominaciones son las dos caras de una misma moneda, ya que el tratamiento del lenguaje natural se realiza aplicando tanto la lingüística como la informática.

En el primer ámbito se desarrollan heramientas centradas en el análisis lingüístico, como son el tratamiento de los corpus, lexicones, y la definición de reglas gramáticas.

Por cuestiones de simplicidad, se denominará indistintamente PLN a las aportaciones de estas dos disciplinas a lo largo de este documento.

Aportaciones de la lingüística y la informática

Cada una de las dos áreas aporta una variedad de herramientas y recursos que son necesarios para la evolución del PLN.

Por un lado, la Informática proporciona lenguajes de programación (por ejemplo PROLOG o LISP) y programas informáticos, aporta soluciones técnicas para el almacenamiento y la estructuración de la información, su representación formal, y también la implantación y optimización de algoritmos.

A su vez, la Lingüística aporta toda la información relativa a la fonética y fonología, gramática, morfosintaxis, semántica, lexicografía y pragmática. De ello se extraen los modelos computacionales del lenguaje, los formalismos de la gramática, y conceptos básicos para la representación del conocimiento.

Dentro del PLN hay que distinguir entre las corrientes que se centran en un tratamiento cuantitativo y las que se centran en el cualitativo.

Las primeras abordan el problema aplicando la Teoría de la Información, las cadenas de Markov, los modelos probabilísticos (como por ejemplo la ley de Zipf), técnicas estadísticas (como la clasificación y el *Clustering*) la Inteligencia Artificial, y las tecnologías del reconocimiento de la voz.

En el caso del análisis cualitativo, el tratamiento se centró en la organización de estructuras conceptuales que permitieran abordar la estructura del lenguaje, tanto en las palabras, lemas, afijos y significados, como en la estructura sintáctica, la morfología, la semántica, y también las semejanzas y diferencias entre diferentes lenguajes.

Esto ha derivado en el desarrollo de varios recursos lingüísticos que han permitido establecer reglas de carácter cualitativo y *a priori* por un lado, y de carácter empírico *a posteriori* por otro.

De los elementos más significativos a nivel teórico, se puede hablar del estudio de la gramática en general, mientras que en el ámbito empírico podemos subrayar el desarrollo de los corpus lingüísticos, y los lexicones.

La diferencia entre los sistemas cualitativos y cuantitativos en PLN son esencialmente las mismas que la aplicación de estos criterios en cualquier área:

> Los sistemas cualitativos se centran en analizar pequeños conjuntos de información profundizando en sus facetas para detectar reglas, criterios y factores determinantes, mientras que el análisis cuantitativo trata de analizar grandes conjuntos de información para extraer patrones mediante métodos estadísticos y numéricos.

El gran potencial de los sistemas cualitativos ha sido que al optar por la exhaustividad, los resultados son significativamente mejores en dominios[1] restringidos, aunque no pueden ser aplicados en dominios más amplios debido al coste de desarrollo que ello supondría. En cambio, los sistemas cuantitativos de la actualidad no plantean grandes problemas de dominio (a pesar de que no hay sistema cuantitativo infalible), aunque su tratamiento es más tosco y sujeto a los criterios descontextualizados de los métodos cuantitativos en general.

A pesar que en la diferentes fases de evolución del PLN desde los años 50 (primeras pruebas de Traducción Automática en Estados Unidos) han habido diferencias entre la visión cualitativa y cuantitativa del PLN, en la actualidad, y después de un gran auge de los sistemas cuantitativos y la toma de conciencia de sus limitaciones, se da un alto grado de equilibrio entre los dos ámbitos.

Las industrias de la lengua

Los inicios del PLN se pueden situar después de la segunda guerra mundial. En esa época, sólo EE.UU. estaba capacitada para llevar a cabo investigaciones de este nivel. Desde entonces hasta ahora, las investigaciones se han ido trasladando también a Europa, y se ha evolucionado desde sistemas muy toscos de traducción hasta algoritmos y software con gran potencia de resolución.

Después de todos estos años de investigaciones, desarrollos y experiencias en entornos controlados, en la actualidad han ido surgiendo una serie actividades con ánimo de lucro

1 Se puede entender por dominio el área de conocimiento que se trata, es decir, el conjunto de elementos que deben tenerse en cuenta para realizar un tratamiento adecuado.

que demandan un *software* que les permita ofrecer soluciones comerciales y servicios asociados para actividades relacionadas con el tratamiento del lenguaje.

Ese tratamiento del lenguaje, que puede ser automatizado o semiautomatizado (combinación de tratamiento por personas y máquinas), ha generado un mercado que buscan una rentabilidad que hasta hace poco no presentaba el PLN.

Es por eso que ha habido un interés creciente en el desarrollo de soluciones ofertadas tanto por instituciones públicas que ofrecen sistemas para la difusión de la lengua, como por entidades privadas que ofertan los servicios mencionados con ánimo de lucro.

A pesar de estos desarrollos productivos, y dado que el PLN es un área en continuo desarrollo, las universidades continúan aportando resultados de investigaciones que permiten mejoras en las técnicas y los algoritmos de tratamiento del lenguaje.

Estas contribuciones junto con las de grandes corporaciones, son las que actualmente lideran el estado actual del PLN. Ello se traduce en que existe una gran cantidad de referencias indicadas en este documento que apuntan a sitios corporativos y departamentos de investigación en universidades o similares.

Dado del carácter complejo del PLN, las inversiones necesarias para desarrollar recursos y algoritmos siguen siendo asumibles sólo por estados o grandes corporaciones.

Recursos lingüísticos

El por qué de los recursos

En PLN, el proceso de información que realiza un ordenador es una tarea mecánica, basada en reglas de control y cálculos.

A lo largo de esa tarea un sistema informático precisa de unos elementos que le permitan las unidades mínimas de tratamiento del lenguaje, en base a la morfología del lenguaje (lemas y flexiones), la estructura y relaciones entre términos (sintaxis) y los significados implícitos (semántica).

Dado que el tratamiento puede realizarse sobre un conjunto relativamente pequeño de términos, el sistema precisa de una base general que contenga todos o gran parte de los elementos que necesite identificar.

En el fondo, lo que aporta la lingüística a la informática y la matemática subyacente es la traducción de los códigos del lenguaje a un sistema cuantificable, y por ello tratable por una máquina mediante algoritmos y reglas de decisión prescritas.

A priori, una frase introducida en un programa de PLN no es más que una secuencia de caracteres. Sobre esta secuencia es necesario que actúen las siguientes capas de proceso:

- Separar las palabras que conforman la secuencia de caracteres (tokenizing).
- Identificar las raíces de las palabras, contrastándolas con lemas y afijos (stemming).
- Analizar la sintaxis de los conjuntos de términos que conforman la frase (analizadores sintácticos).
- Tratar de desambiguar el significado de los términos (en inglés, *Word Sense Disambiguation*, o *WSD*).

En cada una de estas capas es necesario disponer de una información empírica (que podemos denominar los recursos lingüísticos) que posteriormente se contrasta contra la información introducida, y se procesa en base a las reglas del idioma procesado.

Para desarrollar el tratamiento el lenguaje, ha sido necesario que el PLN reciba una serie de recursos de la Lingüística que se describen brevemente a continuación.

Lexicones

Una de las formas más familiares de lexicón son los diccionarios. En esencia, un Lexicón es una colección de términos utilizados en un idioma, sobre los que se puede incluir o no una descripción.

Los lexicones también pueden incluir información sobre prefijos y sufijos, raíces y otras formas o variaciones.

Según el grado de especialización, un lexicón puede ser general de la lengua o bien más especializado. También se puede hablar de lexicones enciclopédicos (Wikipedia podría considerarse como tal) cuando la base de lemas no sólo se refiere a términos del lenguaje, sino a personalidades, aspectos históricos y demás.

El lexicón aporta principalmente información semántica, lo que permite tratar la sinonimia y la antonimia (la relación que establecen dos un significantes en relación a un significado) y la polisemia (la relación entre un significante y dos o más significados).

> Es importante tener en cuenta que en un lexicón, **un lema no siempre equivale a un término**. Existe cantidad de expresiones (empezando por nombres de personalidades célebres) que están formadas por dos términos.

En el momento de trasladar la estructura del lexicón a la informática, es frecuente que aparezcan abstracciones típicas de las estructuras de datos, como son árboles y grafos en general.

También se puede hablar de ontologías (estructuras de representación del conocimiento que aplican el criterio de relaciones entre términos en su sentido más amplio).

Debido al gran esfuerzo que presenta desarrollar y mantener un Lexicón, se han planteado diferentes especificaciones de estructuración y etiquetaje que permitieran la compatibilidad entre sistemas informáticos.

Entre la variedad de especificaciones que existen, se pueden destacar[2] MULTILEX, GENELEX, COMLEX.

También se pueden poner como ejemplos de lenguajes de representación del conocimiento al mismo COMLEX[3], aunque en este caso cabe destacar WORDNET[4], EUROWORDNET[5], y ACQUILEX[6].

Como propuesta de estandarización de la codificación léxica, tenemos el proyecto EAGLES[7].

Por desgracia, parte de estos recursos están restringidos por licencias para uso personal (en esencia no-comercial).

---

2 http://stp.ling.uu.se/~joerg/diplom/node5.html#SECTION00510000000000000000
3 http://nlp.cs.nyu.edu/comlex/index.html
4 http://wordnet.princeton.edu/
5 http://www.illc.uva.nl/EuroWordNet/
6 http://www.cl.cam.ac.uk/research/nl/acquilex/
7 http://www.ilc.cnr.it/EAGLES96/rep2/rep2.html

Por ejemplo, WordNet[8] es de acceso y descarga libre, mientras que EuroWordnet presenta restricciones de copyright, ofreciendo sólo descargas de muestras.

Estos aspectos son importantes aspectos que diferencian el potencial de desarrollo según la lengua de estudio: mientras que el inglés dispone de un lexicón libre y de gran extensión, el desarrollo de un lexicón multilingüe para los idiomas más hablados en Europa se restringe con licencias de uso.

Teniendo en cuenta la total legitimidad de esta opción, es probable imaginar que los pequeños centros de desarrollo podrán aprovechar Wordnet para sus investigaciones, mientras que EuroWordnet quedará restringido a un ámbito más limitado.

Gramática

Dado que se exponen más adelante los objetivos y los elementos de la gramática, cabe comentar que la relación entre la gramática y los recursos lingüísticos es la plasmación de sus reglas y estructuras en un sistema informático que permita su posterior tratamiento.

El grado y el tipo de análisis que se utiliza al plasmar la gramática en un sistema de PLN se diferencia entre superficial y en profundidad. Esta clasificación debe entenderse en clave de los métodos principales utilizados (cuantitativos o cualitativos).

La aportación de la gramática permite desarrollar analizadores morfológicos, morfosintácticos, sintácticos, y sintáctico-semánticos.

Corpus

El corpus lingüístico se compone de ejemplos prácticos de uso de la lengua. Es decir, se pueden encontrar oraciones, párrafos o fragmentos más extensos que proporcionan al sistema ejemplos correctos de uso.

El corpus no incorpora necesariamente documentos completos, ya que pueden ser frases sueltas que exponen la estructura sintáctica de un lenguaje de forma empírica.

Dado que el corpus es la base a partir del cual se va a extraer información para el tratamiento posterior, es importante que su selección sea estricta, de acuerdo con los objetivos del sistema destinatario.

El objetivo principal del corpus no es ser exhaustivo, sino **representativo**.

Según la forma de presentación de un corpus, podemos hablar de corpus orales (grabaciones sonoras) y escritos (fragmentos de textos). Estas dos tipologías se aplican sobre el reconocimiento de voz y el tratamiento de documentos escritos, respectivamente.

Por otro lado, si consideramos el grado de procesamiento de los corpus, podemos hablar de corpus etiquetados y corpus en bruto. La línea lógica de proceso es disponer inicialmente de un corpus en bruto, que recoge un conjunto representativo de ejemplos y que posteriormente es procesado.

En esencia una función del corpus es ofrecer lo que en Inteligencia Artificial se denomina *información de aprendizaje o entreno*, es decir, casos a partir de los cuales el sistema puede detectar patrones (aspectos comunes) y discriminantes (aspectos diferenciadores), de modo que establece criterios para tratar las posteriores entradas de texto.

8 En Debian está disponible aplicando el comando "apt-get install wordnet"

Es por eso que es tan importante la capacidad de síntesis en la selección de textos para el desarrollo de un corpus: primero, porque la no inclusión de elementos clave o la redundancia son situaciones poco eficientes. Segundo, por el esfuerzo de la persona o el equipo responsable de seleccionar el corpus: lo importante es realizar la tarea necesaria, ni más ni menos.

Como ejemplo de corpus, se puede citar a PAROLE[9]. En su versión holandesa[10] se desarrolló en base a la recopilación de textos de diarios y otras publicaciones.

La lingüística y el PLN

Los niveles de estudio del lenguaje

La lingüística ha diferenciado el estudio del lenguaje en base a niveles o aproximaciones a su análisis.

Así, cuando hablamos de los sonidos del lenguaje hablado y a sus representaciones abstractas (denominadas *fonemas*) se habla del **nivel fonético y fonológico** de la lingüística, respectivamente. Cabe tener en cuenta que este nivel sólo es aplicable al PLN en el caso del reconocimiento de la voz (para el caso de la entrada de datos) y de la síntesis de voz (para la salida de datos), temas que no se tratarán en este documento.

Si habláramos de los mecanismos de formación de la adaptación de los lemas al contexto de uso, y de las unidades mínimas de modificación de forma (*morfemas*) estaríamos hablando de la **morfología**.

El PLN realiza un análisis morfológico de los términos. Del análisis se extraen lexemas (también llamados *monemas independientes*) y morfemas (o *monemas dependientes*) que pueden ser **flexivos** o **derivativos**.

Los morfemas adaptan el lexema al contexto de uso, ya sea añadiendo matices de significado (en el caso de los morfemas derivativos) o marcando relaciones gramaticales (en el caso de los flexivos) con el resto de términos de una oración.

Dentro de esta fase se utilizan los *etiquetadores (o taggers) morfológicos* y los *lematizadores*.

Por otro lado, si habláramos de la forma como se relacionan los conjuntos de palabras en los subconjuntos de una frase (denominados *sintagmas*) o en la frase en general, estaríamos hablando del **nivel sintáctico (o sintaxis)**.

A partir de la identificación de la categoría de los términos, es necesario establecer la relación que se establece entre ellos dentro de una oración. Es decir, es necesario identificar el rol del término dentro de la frase y las dependencias con los otros términos.

Esto se realiza mediante el análisis sintáctico. En inglés se identifica el rol de cada término dentro de esta fase como *Part of Sentence* o en si forma abreviada, *PoS*. También se trata el análisis de los sintagmas, que son los conjuntos de términos que se refieren al sujeto, la acción o modificadores (complementos).

Los dos niveles anteriores se acostumbran a unificar hablando del **nivel morfo-sintáctico**.

A nivel léxico, se puede diferenciar entre la **lexicología** y la **lexicografía**. En el primer caso, se dedica a establecer las relaciones entre términos, mientras que en el segundo se centra en analizar la organización y composición de los diccionarios.

El análisis léxico del PLN se basa en la aplicación de los ya mencionados lexicones sobre

9 http://parole.inl.nl/html/index.html
10 http://parole.inl.nl/html-eng/main_info.html

lemas (no debe confundirse con los lexemas), obtenidos al identificar los lexemas y *canonizarlos* (establecer la forma estándar en la que se mostrarían en un diccionario).

Sobre este análisis, el análisis léxico permite identificar el término o locución (dos o más palabras que actúan como unidad, como por ejemplo un nombre propio de una celebridad).

Estas dos disciplinas combinadas se identifican con el **nivel léxico**.

Finalmente, en el **nivel semántico** se identifica la *semántica*, que como su nombre indica se centra en tratar el significado de los términos. Es decir, establece relaciones entre significados y significantes. Esto implica el tratamiento de la ambigüidad, tanto en casos de sinonimia y antonimia, como de polisemia.

El PLN aplica las herramientas desarroladas en la semántica para la desambiguación de los términos. A menudo es necesario haber procesado el contexto para identificar cual de los distintos significados tiene el término. Es por ello que el análisis semántico se realiza después y no antes del análisis sintáctico.

Dado que el el contenido de este documento en relación a la lingüística es exponer su relación directa con el PLN, es de esperar que la información indicada al respecto no sea nada completa.

Pragmática: los valores y el contexto

Antes de entrar en los detalles del análisis del lenguaje, cabe comentar un área de la lingüística que afecta directamente al PLN por ser algo por ahora intratable por un sistema automatizado. En este caso, la pragmática es un ejemplo de lo que queda por recorrer en la completa implantación de las disciplinas de la lingüística dentro del PLN.

La prágmática es un área del lenguaje que se relaciona más con la filosofía del lenguaje que con la propia lingüística. Para los efectos de este documento, se puede afirmar que en la pragmática se analizan los criterios *no gramaticales* por los que un término toma un significado u otro en un contexto determinado. En otras palabras, no hay reglas precisas para el tratamiento de los *rasgos pragmáticos* de un texto o sonido del habla.

Al tratarse de criterios gramaticales, lo que se trata son los aspectos culturales como los valores, tradiciones e idearios colectivos, aunque en esencia la pragmática estudia la intención del emisor.

> En contraste con los criterios gramaticales, la pragmática marca el paso entre la **referencia** (relación directa entre significado-significante) y la **inferencia** (relación probable o contextual entre significado y significante, inducidas por el emisor).

Un autor que puede aportar información de valor sobre el buen uso de la pragmática y la necesidad de transmitir lo más claro posible un mensaje (la ambigüidad también causa errores de comprensión entre humanos) es Paul Grice y su principio cooperativo, conocido en base a las *Máximas de Grice*.

Estas máximas son de cantidad (decir lo justo), calidad (no decir algo falso o no comprobado), relación (ir al grano) y de manera (claridad, no-ambigüedad, ajuste a lo necesario y organización).

En esencia resumen, las Máximas de Grice exponen lo necesario para que la inferencia (interpretación del receptor en base a las intenciones del emisor) se reduzca al mínimo y por lo tanto un sistema automatizado no deba tener en cuenta el contexto del emisor del mensaje.

Como lectura adicional sobre este tema se propone **[Wilson, 2004]**. Trata sobre la teoría

de la relevancia en la comunicación, presentando de forma subyacente las ideas que proponía Grice en sus máximas. Además, dado el tema que se trata, es una lectura de interés para los temas relacionados con la recuperación de la información.

Análisis del lenguaje

Introducción

Se puede entender el análisis del lenguaje como el proceso de interpretación y determinación de la estructura interna de un texto, con el objetivo de captar el mensaje que quiere transmitir.

Por lo general, el objetivo es conseguir que la interpretación permita realizar una tarea (humana o mecánica) de una forma más comprensible y sintetizada. La reducción de la ambigüedad del texto a analizar es uno de los aspectos clave.

Dados los niveles lingüísticos indicados anteriormente, y las herramientas que se corresponden, en este apartado se describirán con más detalle los elementos que intervienen en el proceso.

Análisis superficial y análisis en profundidad

Un analizador sintáctico opera en base a una gramática que permite determinar la estructura que las categorías se estructuran en sintagmas y éstos forman oraciones.

Si se trata de un texto basado en lenguaje general, no especializado, pueden darse situaciones en las que el sistema no sabe determinar la estructura gramatical o sintáctica, ya que la capacidad del sistema nunca es total.

En los casos que el sistema trata con un lenguaje especializado, el número de elementos entrantes se reduce drásticamente, por lo que se puede realizar un análisis en profundidad.

Es por ello que la superficialidad o profundidad dependen de la *exhaustividad* del dominio del lenguaje que se quiere abarcar.

La gramática: tipologías

Se debe entender la gramática como un sistema de reglas que delimita la estructuración del mensaje en base a un código para que el receptor pueda decodificar el mensaje e interpretarlo.

Del mismo modo que un juego puede tener combinaciones prácticamente infinitas aún teniendo muchas reglas, el lenguaje tiene esas mismas características: las reglas de la gramática limitan las posibles construcciones del lenguaje, sin que eso limite el hecho que éste sigue siendo potencialmente infinito (o por lo menos computacionalmente inabordable por un sistema informático).

Se pueden diferenciar entre diferentes tipos de gramática:

- **Prescriptiva**: se centra en construcciones estándares basadas en reglas estrictas que, por ejemplo permiten diferenciar grupos sociales y con ello discriminar a quienes dominan las reglas de quienes no las dominan.
- **Descriptiva**: Establece las reglas sin adjudicar juicios de valor a las posibles alternativas. En esencia se adscribe a una comunidad y acepta las variaciones siempre que sean basadas en palabras correctas dentro de esa comunidad. Es especialmente utilizada en la etnografía, para poder describir la tribu o cultura a estudiar a través de su lenguaje.
- **Tradicional**: reglas heredadas por las lenguas occidentales de Grecia y Roma.

- **Funcional**: Se centra en la organización del lenguaje natural en base a tres normas de adecuación: tipológica, pragmática y psicológica.
- **Generativa**: basada en desarrollar frases correctas desde el punto de vista tradicional, a partir de frases incompletas. En esencia, se basa en deducir los elementos que faltan en una frase (porque a nivel de comunicación humana se pueden presuponer) para que sea una frase correcta. Promovida por Noam Chomsky.
- **Formal**: Gramática precisa y muy estructurada, típicamente usada en los lenguajes de programación de ordenadores[11].

Por su lado, Noam Chomsky propone las siguientes jerarquías[12] en base al nivel de acotación y contextualización de los lenguajes:

- Gramáticas de tipo 0 o sin restricciones. También denominadas gramáticas enumerables recursivamente.
- Gramática tipo 1 o sensibles al contexto.
- Gramática tipo 2 o libres de contexto.
- Gramática tipo 3 o de estados finitos[13].

Analizadores

Los analizadores desglosan el contenido de la entrada de texto para identificar términos, y a partir de ahí realizar el análisis morfosintáctico.

El analizador trata de contrastar si la frase introducida es correcta según las reglas gramaticales que les han sido introducidas.

Tipologías de analizadores

Se pueden diferenciar entre analizadores descendientes y ascendientes. Los primeros (denominados en inglés top-down) parte de la descripción gramatical (las reglas teóricas) y trata de encontrarlas en el texto introducido.

En el caso del analizador ascendiente (down-top), se identifican las estructuras encontradas en el texto de entrada y se contrastan con las reglas gramaticales.

Aparte de si el analizador es ascendiente o descendiente, es necesario saber si el analizador trabaja en amplitud o en profundidad (en relación al dominio de la lengua para el que está diseñado), y el orden en que recorre el texto de entrada para realizar el análisis sintáctico.

En relación a la tipología amplitud-profundidad, cabe comentar que un analizador en profundidad se basa en el conocimiento lingüístico y los dominios especializados. Este tipo de analizadores se basan en las técnicas clásicas, basadas en el análisis estricto de la lengua, sin entrar a fondo en los algoritmos estadísticos aplicables. Es por ello que normalmente se aplican a dominios restringidos de la lengua.

Por otro lado, el análisis superficial se basa en las técnicas de análisis robusto, que se basan en el conocimiento lingüístico, estadístico y tratable mediante algoritmos. La

11 Llama la atención la cercana relación entre la gramática formal y la prescriptiva, y sorprende ver las posibles razones por la competencia que existe entre programadores de diferentes lenguajes de programación.

12 Se puede ver con más detalles técnicos en: http://es.wikipedia.org/wiki/Jerarqu%C3%ADa_de_Chomsky

13 Esta última puede ser tratada por un autómata de estados finitos.

técnica robusta (*superficial*) permite aplicar el análisis a gran variedad de contextos, aunque a cambio de sacrificar en la profundidad del lenguaje y por ello asumiendo errores semánticos, ya que su eje principal son los sintagmas.

Eso implica que el análisis robusto llegará hasta el nivel que su base empírica se lo permita, mientras que en el modelo del análisis profundo, se analiza por completo o no se analiza.

El chunking en el análisis robusto

Como se comentaba, el análisis robusto se basa en el análisis de fragmentos de la frase. El análisis de fragmentos (o análisis fragmental) es denominado en la literatura inglesa como *chunking*.

Esta técnica analiza partes de una frase que el sistema puede procesar fácilmente. Si un bloque no es procesable, queda pendiente sin que el resto de la oración se resienta. Esto permite realizar toda o parte de la tarea, solicitando opcionalmente que el usuario aporte soluciones a lo que ha quedado pendiente.

Para ello, el chunking determina los posibles bloques [*chunks*] a realizar dentro de una oración (por lo general se aplica sobre sintagmas) y asigna los términos a cada bloque.

Al independizar cada una de las partes de la oración, el proceso estadístico se vuelve más robusto porque se delimita las posibles fuentes de ruido (un gran número de términos añade un gran número de combinaciones posibles a analizar) aprovechando la propia estructura superficial (léase *estadísticamente tratable*) del lenguaje.

Funcionamiento del analizador morfosintáctico

Se encarga de determinar la estructura de los términos de la frase a partir de lexicones (repertorios léxicos). A partir de ese contraste, se trata de identificar la categoría morfológica del término (adjetivo, adverbio, sustantivo,...).

Así, en la primera fase, denominada de segmentación [en inglés, *tokenization*], el analizador se encarga de identificar las posibles categorías de cada término. Durante la segmentación se determinan las unidades léxicas, se separan las frases del texto para ser tratadas por partes y en caso que exista, se determinan los nombres propios.

En una segunda fase se realiza la *desambiguación morfológica*, que permite determinar cual de las categorías candidatas es la adecuada para el contexto de la frase. La asignación de la categoría a cada término se denomina *etiquetado* (en inglés *morphological tagger*).

Para ello, se separan las raíces de los morfemas en los términos de las frases (en inglés esa tarea se denomina *stemming*), y se contrasta mediante un diccionario de raíces (lexemas) y afijos (morfemas que pueden ser tanto **prefijos** como **sufijos**).

Finalmente, se trata de concretar todos aquellos condicionantes del contexto mediante el *análisis morfosintáctico*. Eso puede ser aplicado disponiendo de las reglas denominadas *modelo del lenguaje* (que por ejemplo se han extraído del corpus) y la aplicación de los algoritmos de desambiguación que correspondan.

El análisis sintáctico

La base del análisis sintáctico es identificar la estructura interna de una oración. La estructura interna se desglosa en sintagmas, que a su vez vienen constituidos por sujeto, predicado, objetos directo-indirecto, etc.

Se pueden identificar tres tipos de gramáticas según la forma de proceder en el análisis

sintáctico:

- Gramática de constituyentes.
- Gramática funcional.
- Gramática de restricciones.

Para poder realizar el análisis sintáctico es necesario disponer de los siguientes factores:

- Contenidos base:
  - Gramática (morfología, reglas, sintaxis,...)
  - Lexicón (lemas).
- Reglas y procedimientos:
  - Analizador sintáctico.

Malaga: analizador GPL

Se trata de un analizador inicialmente desarrollado por Bjoern Beutel (http://home.arcor.de/bjoern-beutel/) y cuya página está en la página personal del mismo autor (http://home.arcor.de/bjoern-beutel/malaga/).

Malaga permite el análisis morfológico y sintáctico de textos de entrada, y se basa en lo que el autor denomina *Left-Associative Grammars*, desarrolladas por Roland Hausser (http://www.linguistik.uni-erlangen.de/~rrh).

Malaga dispone de un entorno gráfico basado en GTK+ 2.0 llamado malshow, que permite ver tanto el análisis sintáctico de una frase, como la *PoS* de un término y su descomposición (lexema, prefijos y sufijos).

En la página del autor se pueden descargar estructuras gramaticales. Aunque no se encuentra un enlace hacia una gramática en español, se puede acceder a una tesis doctoral que describe su potencial contenido.

Por desgracia, aparte de la resolución de errores, el programa malaga no será mejorado ni actualizado debido a la falta de tiempo de su autor.

Lingüística del corpus

El corpus es uno de los componentes del PLN que más han ganado en importancia a lo largo del tiempo.

Las principales razones para que esto haya sucedido son la introducción de los aspectos empíricos del tratamiento del lenguaje, y el desarrollo tecnológico (que con la introducción de los ordenadores ha permitido procesar cantidades ingentes de información), entre otros.

Otra de las razones es más conceptual, y estriba en que el corpus sintetiza las dos corrientes de análisis de la lingüística: por un lado

Ante esta situación el funcionamiento de un corpus es una combinación de reglas (un corpus son ejemplos de uso y por ello incluye implícitamente las reglas del lenguaje a procesar) y léxico (al tratarse de frases, el sistema puede ser capaz de analizar morfológicamente los términos para poder determinar la estructura sintagmática de las frases).

Un corpus puede ser recopilado a base de oraciones (por lo general inconexas) de una o más lenguas. En este sentido, se diferencian los corpus según el número de lenguas que incluyen, los objetivos para los que se realizan (por ejemplo si son especializados o

generalistas) y el tipo contenido (pueden ser corpus textuales u orales-sonoros).

El diseño de un corpus pretende mostrar en un conjunto controlado, representativo, y relativamente limitado, las reglas de un lenguaje.

Es por ello que depende del corpus y su exhaustividad en el caso parametrizar su contenido. Los criterios aplicables pueden ser de tipo cronológico (por ejemplo, para determinar la variación de la lengua a lo largo del tiempo), de dominio (según la especialidad que traten), canal utilizado para la comunicación, y otros aspectos (factores sociodemográficos del autor, por poner un caso).

También existen los corpus etiquetados, que dan información explícita sobre el contenido. Esto puede implicar información semántica, morfológica o cualquiera que se considere oportuna. En esencia, se trata de indicar información sobre la interpretación que se da a los términos.

El etiquetado condiciona los usos futuros, ya que facilitan al sistema información para el aprendizaje de las reglas y los contenidos de un lenguaje.

Se pueden establecer los siguientes niveles de etiquetado y anotación en un corpus:

- Información bibliográfica.
- Datos estructurales del texto.
- Unidades léxicas.
- Aspectos morfosintácticos.
- Elementos sintácticos.
- Datos semánticos.
- Información discursiva.

Las anotaciones o etiquetados pueden ser realizada/os de forma manual, automática o de forma combinada. En cualquier caso, es importante seguir alguna de estas normativas o estándares para posteriormente disponer de una base compatible con otros sistemas:

- Text Encoding Initiative [TEI] Guidelines. Disponible en: http://etext.lib.virginia.edu/standards/tei/teip4/index.html
- Eagles - Corpus Encoding Standard [CES] . El sistema está basado en el estándar SGML. Disponible en: http://www.cs.vassar.edu/CES/
- En general, se puede utilizar también SGML, aunque por su complejidad esta opción deberá ser delegada en sistemas automatizados.

En esencia, los tres niveles de anotación se basan en lenguajes de marcado[14], como el XML o el propio SGML, por lo que es relativamente sencillo para un sistema interpretar y cargar los datos de un corpus etiquetado según estos criterios.

Los posibles formatos y niveles de anotaciones se relacionan bastante directamente con los niveles del análisis lingüístico (Fonético, Gramatical, Semántico...), siempre teniendo en cuenta las variaciones de tratamiento cuando los corpus son multilingües.

A lo largo del tratamiento del corpus debe ser posible *volver atrás* y disponer del corpus original, separar las anotaciones del propio corpus, y en definitiva, disponer de los elementos originales.

También es de interés disponer de información sobre el quién y el cómo se realizaron las anotaciones, ya que esto puede ayudar al mantenimiento de esa información. Esta información también puede ir complementada por datos bibliográficos (de dónde se sacó

14 Aunque la relación es lejana, vale la pena relacionar el etiquetaje de los corpus con la máxima de utilizar el XHTML y el CSS con criterios semánticos.

esta frase o aquel fragmento).

A pesar de todas estas anotaciones, especificaciones y demás funcionalidades, es necesario tener en cuenta que un corpus no es una herramienta infalible: al pensar en el uso del lenguaje, es importante entender que una muestra representativa, por el hecho de no ser exhaustiva, es falible.

Las aplicaciones de los corpus son varias:

- Investigación en el estudio del habla.
- Aplicación en estudios del léxico.
- Estudio de la gramática y la semántica.
- Filología

Cabe entender también el corpus como la herramienta base que debe permitir el desarrollo (partiendo de una misma base) de léxicos y gramáticas.

Aplicaciones del PLN

Existen tantas potenciales aplicaciones del PLN como situaciones en las que un sistema automatizado actúa como intermediario de un usuario humano o con el propio lenguaje.

Dado el aumento progresivo de la intermediación de sistemas informáticos con usuarios, la progresiva aparición de sistemas que incluyan módulos (subsistemas) de PLN es factible en la medida que:

- Los sistemas sirvan de soporte para:
  - La producción de contenidos con estilo, gramática y ortografía adecuados.
  - Traducción de contenidos.
- Síntesis de información para el soporte a la *extracción de conocimiento*.
- La comunicación directa entre persona y máquina para:
  - Síntesis de la información (*Question Answering* y *Extracción de la Información*). A este respecto puede leerse más verse en **[Vilares Ferro, 2005]**.
  - Interacción directa con aplicaciones (es decir, situar al mismo estilo funcional un botón o un menú de una aplicación y una petición de un usuario, sea escrita o hablada).

Con estas posibles aplicaciones, y las que puedan aparecer en el futuro, se exponen brevemente algunas de las aplicaciones básicas, y se entrará en más detalle para las aplicaciones más potentes y que exponen con más claridad los objetivos y la operativa del PLN.

Verificación ortográfica

Probablemente se trate de una de las aplicaciones más directas y sencillas de realizar, dado que se basa en los niveles más *mecánicos* del análisis lingüístico: el análisis morfológico.

El verificador ortográfico se basa en la disponibilidad de un diccionario que se integra en el software que realice la corrección. En algunos casos no sólo busca si aparece un término determinado, sino que también puede disponer de frecuencias de coaparición de términos en un corpus de lengua. El diccionario detecta términos que no concuerdan y opcionalmente propone alternativas.

En caso que el diccionario no tenga a su disposición algún término que sin embargo exista, es posible incluir una funcionalidad que permita al usuario introducir un nuevo término dentro de un diccionario personalizado.

Las verificaciones ortográficas pueden verse a menudo en los procesadores de texto más habituales. También es habitual en entornos de programación, con bibliotecas de código como ASPELL y PSPELL que proporcionan alternativas o sugerencias en el momento de la escritura.

Verificación gramatical

El análisis sintáctico del texto sube un peldaño en el análisis textual. Ante un texto, el analizador gramatical aplica el reconocimiento de patrones y el análisis sintáctico para detectar los casos que se alejan de los modelos que dispone.

A diferencia del análisis morfológico, el verificador gramatical hace un uso más exhaustivo de técnicas probabilísticas para la identificación de errores gramaticales.

Para el caso del verificador gramatical, el corpus es especialmente importante (aporta los datos base para establecer los patrones de uso).

Verificación de estilo

El objetivo último de los verificadores de estilo es identifcar los elementos que presentan redundancia y densifican el contenido sin aportar valor.

De forma parecida al verificador gramatical, el sistema de verificación de estilo dispone de una serie de modelos predefinidos que establece una serie de rasgos lingüísticos a modo de indicadores (palabras por oración, giros coloquiales, etc.).

Los indicadores anteriores sirven como elemento detector de frases y pasajes sobrecargados.

Como puede verse, la técnica no es equivalente al reconocimiento de patrones aplicado en la verificación gramatical, ya que mientras que el reconocimiento de patrones plantea la estructura como conjunto, la verificación de estilo puede detectar qué parte de la estructura es ineficiente según las reglas de estilo.

Traducción automática

Como su nombre indica, la TA trata de convertir textos de un idioma a otro de forma automática, conservando los aspectos semánticos.

Dependiendo del nivel del lenguaje en el que se base el sistema de traducción y de los recursos lingüísticos utilizados, existen varios sistemas de traducción automática:

- **Traducción directa**: Se encargan de traducir palabra a palabra. Al descartar los aspectos sintácticos, la traducción es muy sencilla y poco útil.
- **Traducción basada en transferencia**: Estos sistemas mejoran considerablemente la calidad final de la traducción, debido a que realizan el análisis morfosintáctico, con lo que realiza una valoración del conjunto y no de las partes separadas de las frases. Estos sistemas son más caros en el desarrollo y por lo general son desarrollado por gobiernos (a través de universidades) y grandes empresas (probablemente también en colaboración con universidades).
- **Interlingua**: Este sistema incorpora un cambio cualitativo en el sistema de traducción, ya que se basa en el desarrollo de un lenguaje abstracto no basado en lenguajes reales. El sistema traduce el texto al idioma abstracto y posteriormente lo traduce al

idioma de destino. Esto implicaría la posibilidad de traducir entre dos idiomas, sin importar la distancia entre los dos idiomas (del Euskera al Chino, por ejemplo).

- **Memorias de traducción**: A diferencia de los sistemas anteriores, que se basan principalmente en reglas, las memorias de traducción se basan en una base empírica, recogida en sistemas de traducción asistida. En caso que se presente una ambigüedad, el sistema solicita la resolución al usuario. La respuesta del usuario se almacena en el sistema, con lo que en la siguiente ocasión ya tiene la respuesta para posteriores ocasiones.

Uno de los aspectos a considerar cuando se desarrolla el sistema, es si se basará estrictamente en dos lenguas o en más, y si se espera que el usuario intervenga o no en el proceso de traducción para aclarar ambigüidades.

Apertium: Traducción por transferencia superficial

Apertium es un sistema de **transferencia superficial** (es decir, se basa en la traducción por transferencia, y se centra en métodos superficiales, es decir, estadísticos).

En concreto, la primera versión de Apertium se centro en realizar la traducción entre lenguas emparentadas. Esta restricción permitió el desarrollo del sistema, reduciendo una gran parte de las diferencias que tendrían dos lenguas lejanas.

La versión 1.0 de Apertium incluía soporte para prácticamente todos los idiomas de la península, incluyendo el Portugués. En la versión 2.0 ya se han incluído funcionalidades para la traducción a idiomas más lejanos, como el Inglés y el Euskera, con la mejora de potencia del sistema que ha supuesto.

El sitio web de Apertium se encuentra en http://apertium.sourceforge.net/, y en el apartado de descargas (http://apertium.sourceforge.net/#downloading) se pueden encontrar *papers* sobre el funcionamiento de este sistema, que tiene licencia GPL[15]. El paquete se puede instalar en Debian mediante paquetes, aunque es necesario descargar los paquetes de la versión 2 (catalán-inglés por ejemplo) de la página de Apertium en SourceForge.

Representación del conocimiento

La relación entre el PLN y la representación del conocimiento (RC) se basa esencialmente en la estructuración de los lemas en forma de sistemas como las ontologías.

Según [Crisholm, 1977], la definición del conocimiento es como sigue:

- El individuo *I* tiene conocimiento sobre *c* si y sólo si:
  - *c* es cierto.
  - I acepta c.
  - c es evidente para I.

Lexicografía y RC

Según [Kirakowski, 1988], existen cinco tipos de conocimientos:

- Conocimiento procedural.
- Conocimiento de objeto y hechos.

15 Se puede leer una traducción no oficial de la licencia GPL en: http://gugs.sindominio.net/licencias/lgpl-es.html. La versión oficial en inglés se encuentra en: http://www.gnu.org/copyleft/gpl.html

- Conocimiento de consecuencias.
- Conocimiento de definiciones.
- Metaconocimiento.

De estos cinco conocimientos, [Cámara, 2004] indica que el *conocimiento de definiciones* es el utilizado en la lexicografía. En esencia, esta afirmación puede resumirse afirmando que los lexicones incluyen las definiciones de términos.

Pero las unidades lexicográficas pueden incluir diferentes sentidos de un mismo término, es necesario disponer de información suficiente para la desambiguación, y determinar las diferencias y los posibles usos del término el varios contextos.

Por ello, es posible que cada entrada léxica disponga de las siguientes informaciones:

- A nivel fonético-fonológico:
  - Transcripción fonética.
- A nivel morfo-sintáctico:
  - Categoría gramatical.
  - Modelo flexivo.
  - Estructura argumental.
- A nivel léxico:
  - Lema o término en sí.
- A nivel semántico:
  - Diátesis
  - Información semántica.
  - Relaciones de sinonimia y antonimia (términos diferentes con los que comparten el mismo significado o el opuesto).
  - Equivalencias en otros idiomas.

Disponiendo de toda esta información, el sistema de representación léxica tiene los siguientes objetivos:

- **Expresividad**: Debe permitir la representación del conocimiento del lexicón.
- **Idoneidad** de referencia y de inferencia: Debe tener capacidad para representar la información en sí como en su aplicación.
- **Herencia**: Permite que se deduzcan las cualidades de ciertos términos en base a términos *padres*. Esto simplifica mucho la introducción de datos ya que se simplifican los aspectos comunes.

Para organizar todos los contenidos del lexicón, se diseñan varias propuestas de estructura según su complejidad y exhaustividad.

Sistemas de representación léxica

Los diferentes sistemas de interrogación establecen la organización interna de la información lexicográfica. Independientemente de la estructura utilizada, el sistema pone a disposición del usuario una interficie de consulta que permite extraer información sobre las unidades lexicográficas, sus definiciones, y las relaciones internas.

El sistema permite extraer un gran volumen de información, y dependiendo de la calidad de la información de entrada, permiten una manipulación y tratamiento más

especializados.

Bases de datos

Organizan la información en base al modelo Entidad-Relación. Esto implica que las relaciones entre términos, significados y demás forman redes de relaciones, más que un modelo jerárquico.

Esta última característica supone un pequeño inconveniente, ya que al no disponer de una estructura jerárquica, no se pueden establecer las inferencias que reducirían el volumen de información [por qué no?].

A pesar de ello, es un sistema que permite representar correctamente la mayoría de términos, dado que técnicamente es un sistema maduro, existe gran cantidad de software utilizable para esta finalidad, con la fiabilidad que la madurez comporta.

Modelos textuales

Por lo general, se entiende por modelos textuales a los corpus textuales. Estos corpus pueden ser utilizados para la representación del conocimiento en la medida que se encuentran anotados/etiquetados.

El nivel y volumen de las anotaciones permitirá acotar con mayor o menor grado el nivel de representación del conocimiento.

Bases de datos léxicas

En este tipo de bases de datos se combinan las funcionalidades de los dos modelos anteriores. Se basa en una fuente de información primaria que es un corpus textual marcado, y por el otro utiliza índices interrelacionados que establecen la estructura de relaciones entre los contenidos de la fuente primaria (el corpus textual).

Bases de conocimiento léxicas

Se centra en el mecanismo de herencia (inferencia) entre elementos, con el objetivo de poder establecer relacinoes enter términos y transferir (inferir) características entre ellos.

Sistemas basados en la unificación

La estructura básica de la representación es la estructuración de las características de los términos. La organización jerárquica permite la definición de relaciones por subsunción, combinación, generalización, reescritura y la posibilidad de establecer herencia.

Redes semánticas

Las redes semánticas, esencialmente idénticas a las ontologías, se basan en el establecimiento de relaciones entre términos en forma de red o grafo.

Dos de los grandes proyectos de definición de redes semánticas son:

- Wordnet
- Eurowordnet

Ambas son Bases de Conocimiento Léxicas, cuyos términos se relacionan entre sí en base a relaciones de sinonimias (synsets) y otros tipos de relaciones semánticas.

En estas bases de datos se representan nombres, verbos, adverbios, adjetivos. Su base de información se centra en el vocabulario básico y amplio, más que especializado o

profundo.

En EuroWordnet existen bases de conocimiento léxico los idiomas mayoritarios de la comunidad europea, aunque también están disponibles el catalán y el euskara.

Las aplicaciones de las redes semánticas se dirigen en esencia a todas las potenciales aplicaciones del PLN. Entre ellos, se pueden destacar los sistemas de traducción automática o la extracción y recuperación de la información entre otros.

Interficies en lenguaje natural

Una interfaz (o interficie) se entiende como el espacio o el entorno en el que el usuario puede comunicarse con la máquina, de forma bidireccional.

Por ejemplo, la interficie de un ordenador se puede entender como el conjunto teclado-mouse-pantalla (en el caso de pantalla, se entiende que no sólo referente al dispositivo, sino al contenido que muestra) en el caso más sencillo.

También podría tratarse de micrófono si se dispusiera de un sistema de reconocimiento de voz que permitiera arrancar programas al emitir su nombre.

Una de las posibles aplicaciones del PLN que se están desarrollando con mayor interés en los últimos tiempos es el de la interacción hombre-máquina (Human computer Interaction, o HCI) mediante el lenguaje natural humano. El desarrollo de este campo recibe el nombre de Interficies en Lenguaje Natural (ILN).

El grado de interacción depende de la generalidad del lenguaje a utilizar, el tipo de usuario que lo utilice y la funcionalidad a tratar.

Esto es especialmente importante debido a que un gran número de potenciales usuarios de la tecnología se quedan al margen debido a la curva de aprendizaje que les supone el uso de instrucciones y sintaxis no naturales.

Además, el uso del lenguaje natural permitiría desarrollar enunciados más complejos, con lo que podría ser posible establecer procesos de búsqueda largos mediante la exposición del problema y la petición concreta de soluciones.

A pesar de ello, las ILN presentan las debilidades derivadas del procesamiento del lenguaje natural: el nivel de cobertura es limitado, la ambigüidad sigue siendo un problema, y aparte de todo esto es un sistema cuyo diseño es caro en relación a otros.

Por otro lado, acarrea los problemas típicos de la HCI tradicional: si la interacción se realiza mediante el teclado y el ratón, el sistema es lento, tedioso y no falto de errores. En caso de realizarlo mediante la voz (Voice-recognition), aparecen cuestiones como la ambigüedad sonora, el ruido de fondo que pueda haber, y la dificultad de identificar la puntuación de las frases mediante las inflexiones de voz.

Es por ello que las ILN se plantean en un horizonte cercano como un sistema de comunicación dentro de un dominio del lenguaje restringido y para unos usuarios ocasionales que no necesiten aprender las funcionalidades de interacción con los sistemas tradicionales.

Dado el precio y las dificultades de refinado, se buscan opciones y funcionalidades que permiten su transportabilidad y reutilización en diferentes entornos.

Lo que en esencia realizan las interficies en lenguaje natural es conectar una frase con un significado u objetivo concretos con las estructuras de proceso de un programa informático.

Es decir, las combinaciones de términos se convierten en *botones* que activan funcionalidades del programa.

El proceso de establecer relaciones entre frases y procesos recibe el nombre de **modelización**, y pasan por la comprensión del mensaje, la elaboración de la respuesta (realizando el proceso que corresponda contra el sistema informático) y su posterior generación.

En esencia, el proceso es bastante equivalente al que se realiza en cualquier aplicación informática con interacción del usuario, salvo que los elementos básicos de interacción (en la actualidad los principales son los entornos de ventanas) incluyen un nuevo elemento que es el lenguaje natural.

Se puede leer información adicional, aunque algo anticuada en **[Long, 1994]**.

Recuperación y extracción de la información

La recuperación de la información es un área profundamente estudiada en la biblioteconomía y la documentación, ya que se refiere a los sistemas que permiten a usuarios que no conocen el contenido de los repositorios, **extraer documentos** que puedan resolver sus dudas.

La recuperación de la información se basa en el tratamiento del documento como conjunto y no por partes, y las técnicas utilizadas más habitualmente se centran en sistemas de indexación (manual, automática o semiautomática) que permiten clasificar los documentos y relacionarlos con una serie limitada de conceptos o elementos incluídos en un lenguaje documental (controlado o no).

El proceso de recuperación de la información basa su potencia en el *establecimiento **a priori** del potencial interés de un documento* para determinados términos o conceptos.

Por su lado, la extracción de información trata el documento con la intención de extraer información concreta. Es decir, la extracción de información no trata el documento sino su estructura gramatical (frases, párrafos) para identificar información

Para el caso de la extracción de la información, se establece primero el motivo de consulta (por lo general en forma de pegunta concreta) y se analiza *a posteriori* la capacidad de extraer información que tienen cada uno de los documentos extraídos.

A diferencia de la recuperación, la extracción de la información requiere el PLN, dado que la extracción de información debe ser sintáctica y semánticamente coherente para el usuario: hay que pensar que la extracción de la información implica separar un bloque de texto de su contexto (con todo lo que ello implica), y procesarlo como tal.

Algunas aplicaciones derivadas de la extracción de la información son:

- Clasificación de documentos: Se trata de una versión simplificada de la extracción de la información, en el sentido que se trata de sintetizar al máximo el contenido del documento para luego organizarlo y categorizarlo.
- Resúmenes automáticos: A pesar que los sistemas de indexación aplican la Frecuencia Inversa de Densidad (IDF)[16] como sistema para generar resúmenes automáticos, el PLN probablemente podría aportar más valor a esta técnica.
- Minería de datos (data mining): Se centra en identificar grupos de datos que ocurren con patrones identificados por el sistema, con lo que se pueden establer correlaciones entre ellos [y opcionalmente relaciones causa-efecto, de pertenencia, etc.]. Una técnica derivada de la anterior es el *web mining*, que aprovecha el hipertexto de la web para extraer información de documentos interrelacionados (enlazados).

En lo referente a la recuperación de la información, el PLN aporta una serie de

16 Ver [Larson, 1998] para disponer de una exposición sobre el IDF y otros sistemas de ponderación de documentos basados en la densidad de términos.

herramientas que favorecen a la delimitación del tamaño del índice de términos o a su tratamiento posterior, permitiendo la indexación mediante:

- lemas o raíces: Elimina los sufijos finales, mediante un analizador morfológico.
- Sintagmas: Permite tratar de forma agrupada información sobre autoridades (políticos, actores, personalidades públicas), entre otros aspectos.
- Sentidos (*semas*) o por conceptos dentro de una red semántica, permitiendo la relación entre los diferentes sentidos.

En general también permite procesar la consulta del usuario, de modo que incluso se puede establecer la recuperación multilingüe de la información, lo que permite recuperar documentos que están escritos en un idioma diferente al que busca el usuario.

Esto también acarrea algunos riesgos derivados del uso del PLN: la incorrecta reducción de la ambigüedad, el dominio insuficiente del lenguaje, etc.

Question answering

Los motores de búsqueda (*Search Engines*) tienen la gran capacidad de recuperar grandes volúmenes de documentos. En esencia, lo que utilizan los buscadores son las técnicas y los algoritmos relacionados con la indexación automática y la recuperación de la información.

La posibilidad de combinar la potencia de los buscadores con la extracción de la información concreta en base a una pregunta, es un método llamado *Question Answering*.

Su utilidad radica en las consultas de carácter situacional y con un componente objetivo importante: es decir, el usuario se encuentra en unas circunstancias concretas que requieren una respuesta concreta, y cuya pregunta se presenta con carácter objetivo (poco ambiguo y sin carga metafórica o personal).

Como en el caso de la extracción de la información, el sistema de QA se basa en devolver textos cortos, fruto de la extracción de contenidos de documentos potencialmente relevantes.

Los condicionantes para esta tecnología son que el conjunto de documentos a procesar es potencialmente enorme, mientras que la respuesta puede ser muy reducida y se dispone de poco tiempo para encontrarla. En pocas palabras, se trata de encontrar una aguja en un pajar.

La solución a estos problemas está realizar una primera fase filtro donde el objetivo es reducir la precisión de la pregunta, a partir de ahí realizar una selección de una colección candidata de documentos, y posteriormente realizar la pregunta precisa sobre esa colección candidata.

Los tipos de preguntas que un sistema de QA puede responder tienen las siguientes tipologías:

- Factuales: Qué, Quién, dónde, cuándo, por qué,
- Sí-no,
- Opinión,
- Causa-efecto.
- Lista de elementos que coinciden con una condición.
- Definición de un concepto
- Preguntas contextuales a un tema determinado.

Por su lado, las respuestas posibles presentan las siguientes tipologías:

- Entidades:
  - Personas
  - Lugares
  - Fechas
  - Números
- Frases o parecidos
  - Definiciones
  - Explicación
  - Método

La extracción de la respuesta se basa en un uso intensivo de patrones, que se asocian a cada tipología de pregunta, y que utilizan información léxica o semantica.

Algunos ejemplos de sistemas de QA son:

- **START**: http://www.ai.mit.edu/projects/infolab/globe.html
- **IO**: http://www.ionaut.com:8400/
- **Webclopedia**: http://www.isi.edu/naturallanguage/projects/webclopedia/
- **AskJeeves**: http://www.ask.com
- **LCC**: http://www.languagecomputer.com
- **AnswerBus**: http://www.answerbus.com/

Bibliografía y referencias

Generales

[Vilares Ferro, 2005] Vilares Ferro, J. (2005). Aplicaciones del procesamiento del lenguaje natural en la recuperación de información en español. http://coleweb.dc.fi.udc.es/cole/library/ps/Vil2005a.pdf [consulta: 12/06/2007]

[Larson, 1998] Larson, R.; Hearst, Marti – Term Weighting and Ranking Algorithms – School of Information Management and Systems – Disponible en: http://www2.sims.berkeley.edu/courses/is202/f98/Lecture17/sld001.htm [consulta: 17/6/2007]

Pragmática

Wilson, D; Sperber, D (2004) – *La teoría de la Relevancia* [en línea] – Revista de Investigación Lingüística, vol. VII, pp. 237-286. Disponible en http://www.um.es/dp-lengua-espa/revista/vol7/relevancia.pdf [consulta: 13/6/2007]

Representación del conocimiento

[Crisholm, 1977] Crisholm, R. (1977) *Theory of Knowledge*. Englewood Cliffs, New Jersey: Prentice-Hall.

[Kirakowski, 1998] Kirakowski, J. (1988). Human Computer Interaction: from Voltage to Knowledge, Cratwell-Bratt Brompley 1988 Lashley, K. (1956). "Cerebral Organization and Behavior." En Solomon, H., (ed.) (1975), *The Brain and Human Behavior* , Baltimore, Williams & Wilkins.

[Arano, 2003] Silvia Arano. *La ontología: una zona de interacción entre la Lingüística y la Documentación* [on line]. "Hipertext.net", núm. 2, 2003. http://www.hipertext.net/web/pag220.htm [Consulta: 17/06/2007]. ISSN 1695-5498

[Cámara, 2004] Lidia Cámara de la Fuente. *La representación lingüística del conocimiento y su relevancia en la ingeniería lingüística* [on line]. "Hipertext.net", núm. 2, 2004. http://www.hipertext.net/web/pag224.htm [Consulta: 17/06/2007]. ISSN 1695-5498

Interficies de lenguaje natural

[Odgen, 1996] Odgen, W.C.; Bernick, Ph. - Using Natural Language Interfaces – Computing Research Laboratory – New Mexico State University – Elsevier Science – 1996. Disponible en: http://crl.nmsu.edu/~ogden/Papers/nl.pn.fm.ps

[Ahrenberg, 1996] Ahrenberg, L.; Dahlback, N - Customizing Interaction for Natural Language Interfaces - Linkoping electronic articles - en *computer and information science*, Vol. 1(1996): nr 1. http://www.ep.liu.se/ea/cis/1996/001/ . October 1, 1996. Disponible en: http://www.ep.liu.se/ea/cis/1996/001/cis96001.ps

[Long, 1994] Long, B. - Natural Language as an Interface Style - Dyncamic Graphics Project – *Department of Computer Science: University of Toronto*, 1994. Disponible en: http://www.dgp.utoronto.ca/~byron/papers/nli.html