La utilización de la Web como recurso en el Procesamiento del Lenguaje Natural

Sandra ROGER[1] **Alexander GELBUKH**[2]
sroger@uncoma.edu.ar gelbukh@cic.ipn.mx

[1] Departamento de Informática y Estadística. Universidad Nacional del Comahue.
Buenos Aires 1400, CP 8300, Neuquén, Argentina. FAX: (54)(0299)4490313.
[2] Centro de Investigación en Computación. Instituto Politécnico Nacional.
Av. Juan Dios Batiz s/n esq. Mendizabal, col. Zacatengo, CP 07738, DF, México.

Palabras Claves: Procesamiento en Lenguaje Natural, Corpus, Web

1 Introducción

La finalidad fundamental del Procesamiento en Lenguaje Natural (PLN) es la automatización de los procesos lingüísticos, tales como la comprensión, producción o adquisición de un lenguaje. En las investigaciones en PLN que consideran al Corpus como componente central, provocan una demandan constante de información léxica detallada sobre amplias áreas de vocabulario.

Numerosas investigaciones ([Bri03],[Kil01b],[Kil01a]) relacionadas con el uso de un mayor corpus han mostrado resultados favorables. El hecho de contar con una considerable ganancia en la exactitud de los resultados al incrementar el tamaño de los datos de entrenamiento, sugiere que tiene sentido poner un gran énfasis en la obtención de grandes corporas, y desarrollos de herramientas que permitirán el acceso efectivo y óptimo al uso de tales recursos.

La mayoría de los desarrollos realizados en la década de los 90 utilizan corpus tales como el British National Corpus (BNC), el cual cuenta con alrededor de 100 millones de palabras, o el TREC QA con un poco mas de 1 millón de documentos, entre otros. Estos corpus eran considerados innovadores al utilizar textos de diferentes tipos.

En esta década el interés de los investigadores se centra en la búsqueda de corpus de mayores dimensiones para sus testeos. En la web contamos con una fuente incalculable de textos de todo tipo. Podríamos considerarla como un nuevo medio de ataque en nuestras investigaciones como menciona Adam Kilgarriff en[Kil01b]

"*El corpus del nuevo milenio es la web.*"

El propósito de este trabajo es presentar la motivación de nuestra línea de investigación, metas y desarrollos futuros. Fundamentalmente, nuestro interés estará en poder capturar el potencial de la web: investigaciones realizadas, líneas en las cuales son mejor candidata o se adaptan de una mejor manera, herramientas disponibles o que pueden ser optimizadas para una utilización más adecuada. En la sección 2 se presenta los lineamientos relacionaldos con la utilización de la web como corpora. En la sección 3 mostramos las concluciones y trabajos futuros.

2 La Web como corpus

El corpus es el eje principal en el desarrollo de PLN. No existe un consenso sobre la naturaleza de la información que el corpus debe contener ni, por supuesto, sobre la manera en la que la información debe ser representada. La tarea de construir un corpus completo para una lengua natural no es trivial.

A mediados de los 90, la web fue utilizada comunmente por los investigadores como un recurso de documentos. Sólo algunos investigadores lo utilizaron como un recurso de conocimiento para un sistema de generalización. El propósito para la utilización de artículos encontrados en la web pueden servir para un rango de posibilidades:

- Confrontación de traductores con términos pocos comunes y a través de una ingeniería de búsqueda poder encontrar evidencias de su uso, contextos, vocabulario asociado etc.
- Para la Recuperación de Información
- Explorar el potencial de la web como un recurso de lenguaje corpora para lenguajes donde son provistos pocos recursos electrónicos.
- Uso de la web para generar entradas enciclopédicas
- Recurso de información léxica. Debido a que la web provee un tesauro de instancias contextualizadas de palabras, esto ofrece la oportunidad para la destilación automática de entradas léxicas de evidencias empíricas.
- Para Word Sense Disambiguation

La utilidad de la web presenta una serie de sus ventajas y desventajas y es un gran motivo de investigación en los próximos años. Los corpus existentes en la actualidad son la odisea comparados con la web. Ésta presenta una serie de características no deseables para un corpus tales como encontrar documentos que no contengan texto, información duplicada, documento apuntando a duplicados y enlaces que deberían apuntar a documentos duplicados y no lo son, el hecho de que la web cambia constantemente, la cantidad de documentos que contienen textos en más de un lenguaje o con lenguajes no identificados

Estas desventajas no son indicadores para decir que la web no es útil en la utilización de un corpus. Sin embargo, para un uso adecuado de la web es necesario establecer ciertos parámetros y restricciones. Se deben desarrollar herramientas que permitan la clasificación de páginas web, que nos brinden información de que clase de texto contienen y demás información que nos posibilite el discernimiento acerca de la información que deseamos utilizar en un corpus.

Un sistema de preguntas-respuestas AskMSR, desarrollado por Eric Brill([Bri03]), fue motivado para mejorar la exactitud al incrementar la cantidad de datos usado en el aprendizaje. Basado en lo expuesto anteriormente ha utilizado la web como un gran recurso de datos que le dió un fundamento para su sistema de pregunta-respuesta. Este sistema es un ejemplo de las tareas que obtendrían beneficios si contaran con datos no comentados[1]. No es factible comentar manualmente grandes corporas, ya que insumen un alto costo, en recursos humanos, en tiempo y en dinero.

[1] Del inglés unannotated

Dado el claro beneficio de contar con datos comentados, es deseable aumentar el grado de investigación relacionado al desarrollo de herramientas y algoritmos que nos permitan, de una manera eficiente realizar esta tarea para magnitudes mayores de datos de los que actualmente están disponibles.

Nuestro objetivo es estudiar un recurso como es la web para poder utilizarlo como corpus. Esto implica la posibilidad del desarrollo de herramientas tendientes a brindar un proceso automático para las diferentes tareas enunciadas anteriormente.

3 Conclusiones y Trabajos Futuros

Se ha presentado una de las líneas de investigación que se está desarrollando dentro del marco de un proyecto de investigación de la Universidad Nacional del Comahue. La web es un rico caudal de textos que pueden ser usados para la investigación en el PLN. Si bien esta información debe ser tratada previamente en muchos de los casos, el beneficio que conlleva la utilización de grandes cantidades de datos para la fase de testeo es importante.

Para poder manipular esta gran cantidad de información, es necesario desarrollar herramientas que posibiliten esta tarea de una manera automática y no manual.

La investigación del uso de texto obtenido de la web y la manipulación de grandes volúmenes está en sus albores. No debemos dejar escapar la posibilidad de desarrollar herramientas tendientes a desarrollar e implementar un proceso automático en la construcción de corpus utilizando este recurso que brinda un sin fin de posibilidaes.

Referencias

[Bri03] Eric Brill. Processing natural language without natural language processing. *Computational Lingistics and Intelligent Text Processing. International Conference, CICLing 2003. México City.*, pages 362–371, 2003.

[Kil99] Adam Kilgarriff. Generative lexicon meets corpus data: the case of non-standard word uses. *In The Language of Word Meaning.*, pages 312–328, 1999.

[Kil01a] Adam Kilgarriff. Comparing corpora. *International Journal of Corpus Linguistics .*, 2001.

[Kil01b] Adam Kilgarriff. Web as corpus. *Proceedings of Corpus Linguistics 2001, Lancaster, March.*, 2001.