Enriquecimiento de Textos en Español Mediante Generación Automática de Hipertexto

Mariano Felice[1]; Fernando R.A. Bordignon y Gabriel H. Tolosa
marianofelice@yahoo.com; {bordi, tolosoft}@unlu.edu.ar

Universidad Nacional de Luján
Departamento de Ciencias Básicas
Laboratorio de Redes

Resumen

Se presenta un proyecto actualmente en desarrollo cuyo objetivo es la creación de un modelo de enriquecimiento de textos basado en la integración de recursos disponibles en el espacio web. El modelo propuesto pretende transformar textos planos lineales en hipertextos que provean información y recursos multimedia sobre entidades reconocidas. Con esta aplicación los usuarios podrán transformar textos en hipertextos auto-explicativos que les posibilitarán una mayor comprensión y les ahorrarán realizar búsquedas individuales de información afín. La evolución al concepto de web 2.0 y la proliferación y popularización de buscadores alternativos, blogs, wikis, servicios de tagging, de *question/answering*, etc. resultan ideales para explotar de manera eficiente los recursos que provee Internet y utilizarlos estratégicamente en el enriquecimiento de texto.

Palabras clave: hipertexto, reconocimiento de entidades, enriquecimiento de texto, *content augmentation*.

1. Introducción

El hipertexto ha revolucionado la forma en que las personas leen un texto, haciéndolo mucho más ágil, interactivo y dinámico dado que brinda la posibilidad de navegar por su contenido de manera no lineal e incluso estar vinculado a otros recursos que lo complementan, por ejemplo mediante contenido multimedia. La existencia de los hipertextos presenta entonces un gran beneficio para los lectores ya que su lectura se ve facilitada además de enriquecida. Por tanto, la preferencia de hipertextos por sobre textos lineales se torna evidente y deseable aunque su disponibilidad está limitada por la creación explícita de sus autores humanos.

Si bien es fácil encontrar hipertextos sobre prácticamente cualquier tema en el espacio web, no siempre se hallan hipertextos que contengan texto o términos exactos deseados por el usuario, o quizás tampoco puedan hallarse hipertextos que contengan vínculos suficientes o interesantes. Por este motivo, sería deseable poder convertir textos particulares de los usuarios en hipertextos, enriqueciéndolos con una variedad de recursos. De esta manera se lograría obtener hipertextos personales automáticos que evitarían su creación manual o las tediosas búsquedas de recursos que complementen los textos leídos.

Los trabajos realizados hasta la actualidad en el área de generación automática de hipertexto han perseguido dos objetivos principales: la generación automática de hipervínculos para navegar la estructura de un documento (por ejemplo, creando índices o referencias cruzadas dentro del documento) y, por otro lado, la vinculación semántica de textos (o parte de ellos) dentro de una

1 Actualmente, se encuentra desarrollando su trabajo final de la Licenciatura en Sistemas de Información, UNLu.

colección finita y estática de documentos. Entre los trabajos más representativos orientados a la creación de hipervínculos estructurales se encuentra el pionero de Frisse [Frisse, 1988], que utilizó un manual de medicina para su experimento, y posteriormente otros proyectos como REXX [Leggett, 1988] y la conversión del Oxford English Dictionary [Raymond, 1988] en formato de hipertexto. Más adelante Fahmy [Fahmy, 1990] y Fuller [Fuller, 1993] realizaron tareas similares aprovechándose de lenguajes estructurados como SGML y XML.

Los desafíos más interesantes han radicado en la construcción de vínculos semánticos entre textos. Diversos autores han abordado este problema utilizando diferentes técnicas y aplicándolas a distintas tareas, tales como ayuda a la navegación, escritura de hipertexto, recuperación de información, etc. Uno de los primeros trabajos corresponde al sistema HieNet [Chang, 1993] que permitía crear hipertextos en forma manual y automática entre los documentos de una colección. Posteriormente, Smeaton et al [Smeaton, 1995] realizaron diversos experimentos mediante los cuales se intentaba vincular documentos de especificación de software de acuerdo al contenido que trataban y cuidando que la red de hipervínculos generada no sea compleja ni confusa. Quizás los aportes más importantes en este área se deban a Allan [Allan, 1995], quien proponía vincular párrafos de textos utilizando un modelo vectorial y asignarles un tipo de vínculo que explicaba la relación existente entre las partes relacionadas.

Entre las investigaciones posteriores más destacadas se encuentran las de Cleary et al [Cleary, 1996], quienes proponen varias técnicas de vinculación que requieren información provista por humanos, Green [Green, 1997], que introduce el concepto de *lexical chaining* para medir la similitud, Jones [Jones, 1998], que creó un sistema editor de hipertexto que sugería vínculos al usuario mientras escribía, y finalmente el trabajo de El-Beltagy et al [El-Beltagy, 2001], quienes crearon un middleware entre servidores web y los navegadores de los usuarios que agrega hipervínculos a términos clave de las páginas visitadas. Si bien este último trabajo es ligeramente similar a la propuesta expuesta en este documento, es aplicable sólo a páginas web y por tal motivo la información que requiere para su funcionamiento se basa en cuestiones esenciales de la navegación, como ser páginas previamente visitadas, URL favoritas, etc.

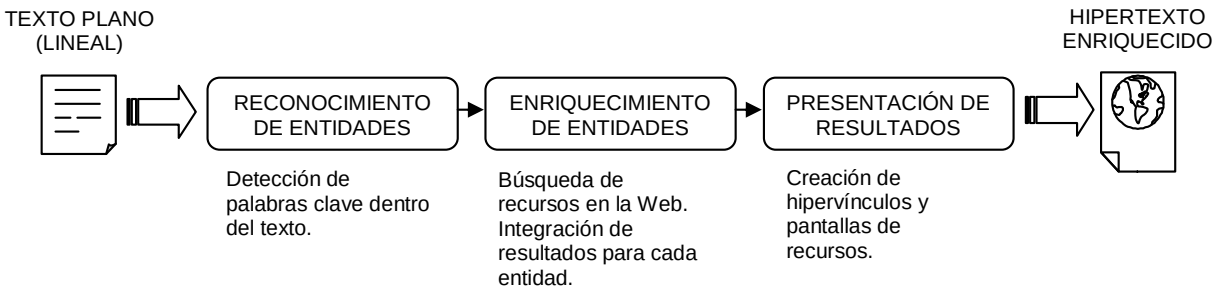
En este proyecto se propone un sistema que basará sus prestaciones en los diversos servicios de información existentes en la web, que en los últimos años han experimentado un gran aumento en su disponibilidad, diversidad, confiabilidad y funcionalidad. La evolución al concepto de web 2.0 y la proliferación y popularización de buscadores alternativos, blogs, wikis, servicios de tagging, de *question/answering*, etc. resultan ideales para explotar de manera eficiente los recursos que provee Internet y utilizarlos estratégicamente en el enriquecimiento de texto.

Este trabajo apuesta a ayudar a los estudiantes en la compresión de textos escolares o informativos, informar a los lectores sobre términos, personajes o acontecimientos desconocidos presentes en artículos de actualidad, servir como buscador contextual compacto de recursos multimedia y actuar como herramienta informativa en otros contextos.

2. Propuesta y Objetivos

El objetivo principal es investigar, diseñar y desarrollar un modelo de aplicación que reciba un texto como entrada y genere una versión enriquecida en formato de hipertexto. El enriquecimiento propuesto consistirá en determinar palabras claves o elementos importantes dentro del texto y vincularlos con información adicional y/o recursos actualizados disponibles en la Web con el propósito de aumentar las capacidades informativas del texto original. Las palabras clave no son

aquellas que resumen el tema del texto ni lo representan sino términos que, al ser complementados con más información, ayudan a comprender mejor el texto o simplemente enriquecerlo. Esta conversión de textos resulta más beneficiosa si se aplica a textos informativos o explicativos, como pueden ser noticias, artículos de revistas, descripciones de personajes, textos históricos, reseñas, etc. Por tal motivo, el diseño de esta aplicación incluirá heurísticas o particularidades que apunten a maximizar el desempeño en este tipo de textos, aunque eso no excluye la posibilidad de que también resulte útil con otros.



Esquema del modelo de enriquecimiento de textos propuesto.

Los recursos que se utilizarán para el enriquecimiento serán tomados de la World Wide Web, dado que es la fuente de información digital más completa, variada y actualizada que existe en la actualidad. Además, la gran disponibilidad de recursos en la Web otorga al modelo tolerancia a fallos (al recurrir a otros recursos), completitud (dada la variedad de contenidos), actualización permanente y objetividad (al integrar contenidos de múltiples fuentes). Inicialmente, se apuntará a enriquecer un número limitado de entidades del texto de entrada, posiblemente lugares geográficos, organizaciones, nombres comerciales y personajes, aunque no se descarta la posibilidad de incluir nuevos reconocimientos en el futuro.

Texto original

Texto enriquecido

3. Metodología y Problemas a Resolver

A partir de la propuesta del modelo de conversión de texto plano lineal en hipertexto mediante el enriquecimiento con hipervínculos se trabajará en las tres partes fundamentales, las cuales presentan desafíos particulares:

- Reconocimiento de Entidades:
  - Detectar las palabras clave a enriquecer. Para resolver esta cuestión será necesario investigar y adaptar técnicas de *Named Entity Recognition*. Entre las más útiles y significativas se encuentran trabajos como [Carreras, 2002], [Cucerzan, 1999], [Magnini, 2002], [Maynard, 2001], [Toral, 2005] y [Toral, 2006].
  - Efectuar una desambigüedad de los términos elegidos para evitar errores semánticos. Esto implicará, entre otras cosas, detectar contextos en varios niveles dentro del documento, posiblemente a nivel oración, párrafo y texto completo.

- Enriquecimiento de Entidades
  - Definir los tipos de recursos a vincular y las fuentes de donde se obtendrán, lo que implica identificar los servicios web a utilizar.
  - Obtener vínculos a recursos mediante la utilización de los servicios seleccionados en la etapa anterior. Para realizar este objetivo resultará necesario definir una estrategia de búsqueda, como ser la creación de *queries* automáticos, la búsqueda en listas u ontologías, la consulta de enciclopedias on-line, etc. Los trabajos de Janevski y Dimitrova [Janevski, 2002] y Dowman et al [Dowman, 2005] para el enriquecimiento de video son ejemplos interesantes de cómo puede lograrse tal extracción de información.

- Presentación de Resultados
  - Diseñar una interfaz apropiada para la presentación del hipertexto generado que facilite la lectura, la presentación de los recursos insertados y su navegación.

En todos los casos, resultará necesario integrar y adaptar técnicas del área de Recuperación de Información y de búsquedas en la web que permitan seleccionar en forma eficiente y apropiada los recursos que serán vinculados.

4. Avance del Proyecto

Como resultado del proyecto propuesto, se espera alcanzar las siguientes metas, algunas de las cuales ya han sido concretadas:

- Seleccionar o definir un método de reconocimiento de entidades para determinar los términos candidatos a ser enriquecidos dentro de los textos. (finalizado)
- Diseñar un modelo de enriquecimiento de textos capaz de integrar una variedad de recursos actualizados disponibles en la World Wide Web que provean información confiable. (finalizado)
- Planificar y definir una estrategia de combinación de recursos óptima cuyos resultados sean una selección de información y recursos complementarios a una entidad del texto (en desarrollo).
- Diseñar un modelo de presentación apropiada de los resultados, lo que implicará determinar la mejor cantidad de entidades a enriquecer, analizar las formas más adecuadas para

presentar la información de enriquecimiento, facilitar la navegación del texto y los recursos integrados, etc. (en desarrollo).
- Implementar el modelo de enriquecimiento como un prototipo de aplicación de web (en desarrollo).
- Realizar evaluaciones internas y externas del rendimiento y la utilidad de la aplicación desarrollada. (a realizar)

5. Referencias

[Allan, 1995] Allan, J. Automatic Hypertext Construction. PhD thesis, Cornell University, 1995.

[Carreras, 2002] Carreras, X.; Márquez, L. y Padró, L. Named Entity Recognition Using AdaBoost. En *Proceedings of the 2002 CoNLL Workshop*, 2002.

[Chang, 1993] Chang, D. T. HieNet: A User-centered Approach for Automatic Link Generation. En *Proceedings of the Fifth ACM Conference on Hypertext (Hypertext'93),* ACM, 1993.

[Cleary, 1996] Cleary, C; Bareiss, R. Practical Methods for Automatically Generating Typed Links, *Proceedings of the seventh ACM conference on Hypertext*, 1996.

[Cucerzan, 1999] Cucerzan, S. y Yarowsky, D. Language Independent Named Entity Recognition Combining Morphological and Contextual Evidence. En *Proceedings of the 1999 Joint SIGDAT Conference on Empirical Methods in NLP and Very Large Corpora*, 1999.

[Dowman, 2005] Dowman, M.; Tablan, V.; Cunningham, H. y Popov, B. Web-Assisted Annotation, Semantic Indexing and Search of Television and Radio News. *14th International World Wide Web Conference*, 2005.

[El-Beltagy, 2001] El-Beltagy, S. R.; Hall, W.; DeRoure, D. y Carr, L. Linking in Context. En *Proceedings of the Twelfth ACM Conference on Hypertext and Hypermedia (Hypertext '01)*, 2001.

[Fahmy, 1990] Fahmy, E. y Barnard, D. T. "Adding Hypertext Links to an Archive of Documents" en *The Canadian Journal of Information Science*, 1990.

[Frisse, 1988] Frisse, M. E. Searching for Information in a Hypertext Medical Handbook. En Communications of the ACM (CACM), 1988.

[Fuller, 1993] Fuller M.; Mackie E.; Sacks-Davis, R. y Wilkinson R. Structured Answers for a Large Structured Document Collection. En *Proceedings of ACM SIGIR '93*, 1993.

[Green, 1997] Green, S. Automatically Generating Hypertext by Computing Semantic Similarity. PhD thesis, University of Toronto, 1997.

[Janevski, 2002] Janevski, A. y Dimitrova, N. Web Information Extraction for Content Augmentation. En *Proceedings of the IEEE International Conference on Multimedia and Expo*, 2002.

[Jones, 1998] Jones, S. Link as you Type: Using Key Phrases for Automated Dynamic Link Generation. Working Paper 98/16, University of Waikato, New Zealand, 1998.

[Leggett, 1988] Leggett, J. J.; Nunn, D.; Boyle, C. y Hicks, D. The REXX Project: A Case Study of Automatic Hypertext Construction. Hypermedia Research Lab, Dept. of Computer Science, Texas A&M University Technical Report TAMU 88-021, 1988.

[Magnini, 2002] Magnini, B.; Negri, M.; Prevete, R. y Tanev, H. A WordNet-Based Approach to Named-Entites Recognition. En *Proceedings of SemaNet02, COLING Workshop on Building and Using Semantic Networks*, 2002.

[Maynard, 2001] Maynard, D.; Tablan, V.; Ursu, C.; Cunningham, H. y Wilks, Y. Named Entity Recognition from Diverse Text Types. En *Recent Advances in Natural Language Processing 2001 Conference*, 2001.

[Raymond, 1988] Raymond, D. R. y Tompa, F. W. Hypertext and the Oxford English Dictionary. *Communications of the ACM*, 1988.

[Smeaton, 1995] Smeaton, A. F.; Morrissey, P. J. Experiments on the Automatic Construction of Hypertexts from Texts. En *New Review of Hypermedia and Multimedia*, 1995.

[Toral, 2005] Toral, A. DRAMNERI: A Free Knowledge Based Tool to Named Entity Recognition. En *Proceedings of the 1st Free Software Technologies Conference*, 2005.

[Toral, 2006] Toral, A. y Muñoz, R. A Proposal to Automatically Build and Maintain Gazetteers for Named Entity Recognition by Using Wikipedia. En *Workshop on New Text, 11th Conference of the European Chapter of the Association for Computational Linguistics*, 2006.