# Evaluación de un lematizador independiente del idioma sobre una colección Inglés-Español

George Zoghaib Chakhtoura
Maestría en Computación
Universidad de Costa Rica
Montes de Oca, San José, Costa Rica

**Resumen:** Se exponen y analizan los resultados obtenidos sobre la comparación de dos lematizadores, uno dirigido al idioma inglés y otro independiente del idioma, implementado mediante el método de variedad de sucesores, sobre una colección que posee documentos en inglés y español, utilizando consultas en inglés.

## 1. Introducción y metodología

Los experimentos realizados hasta el momento, no han logrado confirmar satisfactoriamente las mejoras que la lematización implica en la precisión de los motores de búsqueda. Sin embargo, demuestran que la utilización de los términos lematizados no produce impactos negativos al dicho motor (Baeza Yates, Ribeiro-Neto 1999).

Dado esto, aun se considera que la reducción de las palabras a indexar en un motor búsqueda a sus lemas correspondientes puede ser útil para mejorar el desempeño y la precisión de dicho motor, pues se reducen las diversas variantes de una raíz a un solo concepto.

En general, los lematizadores pueden ser desarrollados para un idioma específico, como lo son los lematizadores de análisis flexivo, o independientes del idioma. Estos últimos pueden ser utilizados en colecciones donde no es posible determinar un idioma dominante sobre las mismas.

La implementación de un lematizador no dirigido a un idioma específico puede ser realizada mediante el algoritmo de variedad de sucesores para cada letra de una palabra. Esto permite agrupar diferentes términos bajo una misma raíz.

El enfoque principal del presente experimento se encuentra dirigido a determinar efectividad de un lematizador no dependiente del idioma en un vocabulario español-inglés, utilizando consultas en inglés, en contraposición con el lematizador flexivo para inglés implementado mediante Snowball (Snowball 2004).

Para realizar dicha evaluación, se ha seleccionado una colección de información turística sobre Costa Rica, la cual contiene 618 documentos en inglés y español.

Adicionalmente, se utilizan tres consultas predeterminadas, para las cuales se conocen los primeros veinte documentos relevantes a las mismas, dado que fueron determinados durante la creación de la colección por un grupo de alrededor de veinte personas. Por medio de este conocimiento, es posible validar la precisión de los documentos retornados por medio del motor de búsqueda, utilizando ambos lematizadores.

Cabe mencionar el hecho de que se utiliza el lematizador de Snowball para el idioma inglés y no para el español, pues las

consultas son realizadas en el primer idioma mencionado.

Con respecto al lematizador independiente del idioma, el mismo es implementado mediante el algoritmo de variedad de sucesores, utilizando el método de corte arbitrario en el cuarto nivel. Este lematizador fue desarrollado, inicialmente, para trabajar sobre los idiomas inglés y español principalmente, por lo que es posible que, en caso de ser necesario su utilización para otros idiomas occidentales, el mismo deba ser modificado.

Para realizar las consultas sobre la colección, se modificó el motor de búsqueda realizado por Ricardo Pezo y George Zoghaib para el curso de Laboratorio de Recuperación de Información, impartido en la Universidad de Costa Rica en el primer período del año 2004.

Dos índices fueron creados; utilizando el lematizador de Snowball para el primero, y el de variedad de sucesores para el segundo. De igual forma, las consultas fueron lematizadas con cada lematizador, y se realizaron utilizando el mismo motor de búsqueda.

## 2. Resultados obtenidos

La evaluación de las consultas realizadas sobre la colección lematizada se basa en los primeros veinte documentos retornados. Los mismos son comparados con los considerados como relevantes, midiendo de esta forma la precisión de los resultados.

Para la primera consulta, *Rafting*, los resultados del motor utilizando el lematizador de Snowball son los siguientes:

- 205 documentos fueron retornados.
- En los primeros 20, 8 son considerados como muy relevantes.
- Entre los primeros 5 documentos retornados, ninguno corresponde a un documento considerado muy relevante.

Por otro lado, los resultados utilizando el lematizador independiente del idioma son los siguientes:

- 205 documentos fueron retornados.
- En los primeros 20, 7 son considerados como muy relavantes.
- El primer documento considerado como relevante, aparece de sexto entre los retornados.

En relación con la segunda consulta, *Surfing*, los resultados utilizando el lematizador de Snowball fueron:

- 111 documentos retornados.
- 15 de los primeros 20 son considerados como relevantes.
- Los primeros siete documentos retornados son relevantes.

En el caso del lematizador independiente del idioma, se obtuvieron los siguientes resultados con respecto a la consulta *Surfing*:

- 134 documentos retornados
- 14 de los primeros 20 documentos son relevantes.
- Los primeros 11 documentos retornados son relevantes.

Finalmente, la consulta *Backpacking* en combinación con el lematizador de Snowball para inglés retornó 166 documentos, de los cuales ninguno de los primeros 20 es relevante.

En el caso de la misma consulta, utilizando el lematizador de variedad de sucesores, los resultados fueron los siguientes:

- 212 documentos retornados.
- 6 de los primeros 20 son relevantes.
- El segundo documento del resultado es el primero en ser considerado como relevante.

## 3. Conclusiones

Analizando los resultados obtenidos utilizando los dos lematizadores, es claro que no es posible determinar la mayor eficacia o precisión de uno sobre otro, pues, exceptuando la tercera consulta, la diferencia entre los primeros veinte documentos relevantes retornados es mínima a favor del lematizador de Snowball.

Con base en esto, es posible afirmar que en una colección de documentos de inglés y español, distribuida normalmente, un lematizador independiente del idioma provee resultados cercanos a un lematizador asociado al idioma inglés, cuando se utilizan consultas en inglés.

Por otro lado, cabe mencionar que es necesario realizar estas consultas con una colección de mayor tamaño para obtener datos más específicos y apegados a la realidad.

Adicionalmente, es claro que debe de realizarse el mismo experimento utilizando consultas en español, para determinar si se mantiene la relación cercana de los resultados entre un lematizador dirigido a un idioma específico – en este caso español – y otro independiente del mismo.

## 4. Bibliografía

1. Baeza-Yates, R. Ribeiro-Neto B., 1999 "Modern Information Retrieval" Addison Wesley. ACM Press, New York.

2. Sitio web de Snowball [http://snowball.tartarus.org] – visitado el 9 de Julio, 2004.

3. Figuerola, C.G.; Alonso Berrocal, J.L.; Zazo Rodríguez, A.F.; Gómez Días Raquel. Mayo 2002. Informe Técnico: Recuperación de información utilizando el modelo vectorial. Participación en el taller CLEF-2001. Departamento de Informática y Automática. Universidad de Salamanca.

4. McNamee, Paul and Mayfield, James. 2002. Single N-gram Stemming. The Johns Hopkins University Applied Physics Laboratory.

5. Figuerola, C.G.; Alonso Berrocal, J.L.; Zazo Rodríguez, A.F.; Gómez Días Raquel. 2001. Stemming in Spanish: A First Approach to its Impact on Information Retrieval. Universidad de Salamanca, Spain.