

Simplificación de texto para facilitar la comprensión lectora del usuario final

Susana Bautista, Pablo Gervás

Departamento Ingeniería del Software e Inteligencia Artificial,
 Facultad de Informática,
 Universidad Complutense de Madrid, Spain
 subautis@fdi.ucm.es, pgervas@sip.ucm.es

Resumen El modo en que se escribe o se presenta la información puede excluir a muchas personas, especialmente a quienes tienen problemas para leer y escribir o para entender. El sistema propuesto en el presente artículo pretende servir de ayuda para la adaptación semiautomática de textos a lectura fácil. Partiendo de un análisis sintáctico del texto, se simplifican sus estructuras sintácticas y léxicas, siguiendo las guías propuestas a nivel europeo y se generan textos adaptados a la lectura fácil que permitan el acceso a la información como derecho fundamental de la persona.

Key words: simplificación, adaptación, análisis, métricas, fácil lectura

1. Introducción

El acceso a la información plantea dificultades a determinadas personas que tienen problemas relacionados con su capacidad de leer y de entender como por ejemplo las personas que padecen retraso mental u otro tipo de diversidad funcional, las personas que solamente han tenido acceso a una formación cultural limitada, las personas con problemas sociales y los inmigrantes cuya lengua materna no es la lengua oficial de su país de adopción. El objetivo de este trabajo es desarrollar una arquitectura capaz de llevar a cabo de manera semiautomática e interactiva la tarea de generar textos en formato de lectura fácil. El sistema servirá de ayuda a quien tenga que redactar versiones adaptadas de textos. Se considera objetivo prioritario el conseguir una generación de textos capaz de satisfacer las directrices de accesibilidad establecidas en los Web Content Accessibility Guidelines 2.0 [1] relativas a la facilidad de comprensión de los contenidos. Las Normas Uniformes sobre la Igualdad de Oportunidades para Personas con Discapacidad, de Naciones Unidas, requieren que se hagan accesible todos los servicios públicos de información y documentación para que sean fácilmente accesibles a todos. [2]

2. Estado del arte

Con objeto de ampliar los esfuerzos llevados a cabo por las Naciones Unidas, la asociación europea ISLMH [3] ha llevado a cabo un proyecto para desarrollar las guías para la elaboración de materiales de lectura fácil, *Directrices Europeas para generar Información de Fácil Lectura*. En estas directrices se recogen las sugerencias y las características que deben tener los textos para que no presenten dificultades de lectura y son las que nosotros seguimos durante el proceso de adaptación para obtener texto de lectura fácil.

Existen diferentes métricas para medir la fácil lectura de un texto, basadas en ecuaciones matemáticas cuyas medidas correlativas son los elementos de escritura, como el número de pronombres personales en el texto, el número de sílabas por palabras o el número de palabras por oración en el texto. Están disponibles para el inglés [4] y para el castellano [5]. No tienen en cuenta el significado de las palabras, ni de las frases, por lo que no sirven para evaluar ni la corrección lingüística ni el contenido informativo del texto.

Dentro del análisis sintáctico se distingue entre el análisis de constituyentes y el análisis de dependencias. El análisis de constituyentes se caracteriza por el uso de la relación de inclusión (unos sintagmas incluyen a otros y, en el caso básico, se tienen sintagmas compuestos por unidades léxicas). El análisis de dependencias se caracteriza por el uso de relaciones binarias (de dependencia) entre unidades léxicas. Para el inglés existen analizadores de dependencias como Minipar [6] y el Stanford Parser [7], que además lleva a cabo análisis de constituyentes. Entre los analizadores para el castellano, destacamos FreeLing [8] y JBeaver [9] que realizan análisis de dependencias, y VISL [10], que realiza análisis de constituyentes.

En la parte correspondiente a la lexicalización a la hora de adaptar un texto, usamos bases de datos léxicas. Entre ellas para el inglés cabe destacar WordNet [11], y para el español, entre otros idiomas europeos, EuroWordNet [12]

3. Propuesta. Conversion Semiautomática

El sistema propuesto pretende servir de ayuda para la adaptación de textos a lectura fácil. Con el sistema se quiere facilitar la simplificación de textos para que sean accesibles a un mayor número de personas. Para ello, se va a llevar a cabo una serie de transformaciones con las que queremos que mejore la comprensión del texto. La evaluación del sistema se realizará de una manera empírica a partir de los textos originales y de las versiones en lectura fácil que se van a generar con el sistema.

El proceso a seguir para llevar a cabo la conversión semiautomática de un texto dado a su versión en lectura fácil es el siguiente: Partimos de la versión original de un texto, al que se le van aplicar las *métricas* correspondientes para medir su nivel de dificultad, para poder hacer luego la comparación con el texto resultado. Para cada frase del texto de entrada, se realiza un *análisis* utilizando uno de los analizadores sintácticos que tenemos como recursos (Tabla

1). Partiendo del resultado del parser, se hace la *conversión* sintáctica usando las reglas de conversión definidas a partir de las Directrices Europeas [3]. Tras la simplificación sintáctica realizada, se aplican las *métricas* al texto resultado para estudiar como afecta la simplificación al nivel de dificultad del texto con el que estamos trabajando. El siguiente paso en el proceso es la lexicalización, para ello se realiza una *consulta léxica* al recurso correspondiente y se aplican las reglas de sustitución léxica definidas en cada caso. Tras este último paso de adaptación, se vuelven a aplicar las *métricas* al texto final para poder realizar un estudio comparativo con el texto del que partíamos y el texto que hemos obtenido como salida del sistema.

En la Tabla 1 presentamos los recursos que nos planteamos utilizar en cada paso.

Recursos	Inglés	Castellano
Análisis Constituyentes	Stanford Parser	VISL
Análisis Dependencias	Minipar, Stanford Parser	JBeaver, FreeLing
Recursos Léxicos	WordNet	EuroWordNet

Tabla 1. Recursos con los que vamos a trabajar

4. Discusión y Conclusiones

Hay que llevar a cabo un estudio comparativo entre el análisis de constituyentes y el análisis de dependencias, dentro de la simplificación sintáctica, para ver con cuál de los dos obtenemos mejores resultados para la simplificación de los textos. Con cuál se consigue una mejora significativa en la comprensión del mismo por parte del usuario final. A partir de la Tabla 1 podemos ver las distintas posibilidades que tenemos de llevar a cabo el proceso planteado para el sistema. Las reglas de conversión, hay que definir las en términos de componentes sintácticos y también en términos de relaciones de dependencias para poder llevar a cabo el proceso de simplificación usando cualquier tipo de analizador.

El sistema ya está en desarrollo, [13] y actualmente trabaja usando las métricas correspondientes para el inglés, utilizando el Stanford Parser como analizador de constituyentes y como recurso léxico utiliza WordNet.

Además de la evaluación realizada a partir de las métricas hay que realizar un estudio exhaustivo para medir transformaciones erróneas que se hayan realizado, identificar que oraciones son gramaticalmente incorrectas, cuáles son gramaticalmente correctas pero semánticamente incorrectas tras el proceso de simplificación. De tal manera que consigamos medir la simplicidad alcanzada frente a la información perdida en el proceso, para poder evaluar el sistema desarrollado.

Nuestra principal apuesta es que el sistema puede funcionar, y puede ser muy útil para distintos colectivos. Las herramientas y recursos que proponemos utilizar existen y están disponibles.

Referencias

1. W3C: Web content accessibility guidelines 2.0. Technical Report WCAG20 (2008)
2. Nations, U.: Normas uniformes sobre la igualdad de oportunidades para las personas con discapacidad. Technical report (1994)
3. Freyhoff, G., Hess, G., Kerr, L., Menzel, E., Tronbacke, B., Veken, K.V.D.: European guidelines for the production of easy-to-read information for people with learning disability. Technical Report ISLMH (1998)
4. Barrow, C.: Readability metrics. <http://w-shadow.com/blog/2009/04/28/calculating-readability-metrics-in-php/> (2009)
5. Moro, P., Cabero, M., Rodríguez, J.L.: Ecuaciones de predicción de lecturabilidad. Technical report, e-spacio uned (1993)
6. Lin, D.: Dependency-based evaluation of minipar. In: In Proc. of Workshop on the Evaluation of Parsing Systems. (1998)
7. Klein, D., Manning, C.D.: Fast exact inference with a factored model for natural language parsing. In: In Advances in Neural Information Processing Systems 15. (2003) 3–10
8. Atserias, J., Casas, B., Comelles, E., González, M., Padró, L., Padró, M.: Freeling 1.3: Syntactic and semantic services in an open-source nlp library. In: Proceedings of the fifth international conference on Language Resources and Evaluation (LREC 2006), ELRA, Genoa, Italy (May 2006)
9. Herrera, J., Gervás, P., Moriano, P., Moreno, A., Romero, L.: JBeaver: un Analizador de Dependencias para el Español Basado en Aprendizaje. In: XII Conferencia de la Asociación Española para la IA. Volume 2., Salamanca, España (11/2007 2007) 211–220
10. Bick, E.: The VISL System: Research and applicative aspects of IT-based learning. In: Proceedings of NoDaLiDa,(Uppsala). (2001)
11. Miller, G.A., Beckwith, R., Fellbaum, C., Gross, D., Miller, K.J.: Introduction to WordNet: An On-line Lexical Database. *Int J Lexicography* **3**(4) (1990) 235–244
12. of Amsterdam, U.: Eurowordnet. <http://www.illc.uva.nl/EuroWordNet/> (1999)
13. Bautista, S., Gervás, P., Madrid, R.: Feasibility Analysis for SemiAutomatic Conversion of Text to Improve Readability. In: The Second International Conference on Information and Communication Technologies and Accessibility. (2009)