Creado el 11/02/2014
Última versión: 11/02/2014
Horacio Saggion, Universidat Pompeu Fabra, Barcelona
Comunidad moocTLH
Esta obra está bajo una licencia de Creative Commons Reconocimiento-CompartirIgual 3.0 Unported.

Retos y oportunidades en las TLH/PANORAMA ACTUAL

Simplificación automática de textos

Hola.

Soy Horacio Saggion, profesor investigador en el departamento de tecnologías de la universidad Pompeu Fabra en Barcelona.

Hoy les presentaré una breve introducción a la simplificación automática de textos.

Los temas que tocaré en esta oportunidad son los siguientes:

- Primero introduciré algunos conceptos básicos de simplificación de textos.
- Seguidamente veremos tres problemas típicos que se estudian en simplificación de textos.
- Luego veremos brevemente sistemas basados en reglas.
- y continuaremos con sistemas que aprenden a simplificar.
- Finalmente discutiré algunos temas relacionados.

Conceptos básicos

La simplificación de textos consiste en transformar un texto en un equivalente que sea más fácil de leer por una audiencia determinada.

Generalmente se transforman las oraciones complejas en oraciones más simples y se reemplazan las palabras difíciles por palabras más fáciles.

Ahora bien. ¿Para qué sirve la simplificación de textos?

Hay muchos colectivos de personas que pueden beneficiarse de la simplificación de textos: por ejemplo las personas semianalfabetas que no han adquirido conocimientos suficientes de la lengua, las personas con discapacidad intelectual que pueden leer textos de complejidad moderada, los afásicos que tienen varias dificultades con el lenguaje, los sordos, los disléxicos, etc.

Aquí vemos un ejemplo de simplificación de textos producida por un profesional. Veamos una oración típica de una noticia periodística. Esta oración puede resultar bastante compleja para algunas personas pues tiene una sintaxis complicada y además un vocabulario poco común como por ejemplo “filtrar cables” o “proporcionar trato inhumano”. Además contiene nombres propios que pueden ser difíciles de entender.

Un simplificador humano produce la siguiente simplificación que consiste de 4 oraciones simples.

texto no simplificado	texto simplificado
Amnistía Internacional acusó a las autoridades estadounidenses de proporcionar un "trato inhumano" a Bradley Manning, un soldado acusado de filtrar “cables” de la diplomacia norteamericana a portal Wikileaks.	Estados Unidos trata muy mal a un soldado detenido. El soldado se llama Bradley Manning. Bradley Manning está detenido por dar información del gobierno de Estados Unido a Wikileaks. Wikileaks es una página web donde se da información sobre asuntos de interés público.

Por un lado una oración con la idea principal de la noticia (sobre cómo se está tratando al soldado). Por otro lado una oración que indica el protagonista de la noticia (quien es el soldado). La tercera oración contiene la información adicional (sobre por qué el soldado está detenido). y finalmente una cuarta oración que introduce una definición de Wikileaks para explicar de qué se trata.

Encontrar ejemplos de textos originales y sus simplificaciones no es fácil. Sin embargo sí que hay sitios Web que publican textos accesibles, tales como la enciclopedia Wikipedia en su versión simple (para inglés).

Y el portal Noticias Fáciles para el español.

Tres problemas en simplificación de textos

Uno de los problemas es identificar la dificultad de un texto a fin de decidir si es apropiado para una audiencia determinada. En general esto significa clasificar el texto según su nivel de legibilidad.

El segundo problema consiste en simplificar la sintaxis de las oraciones. Construcciones como las subordinaciones o las coordinaciones son generalmente simplificadas en oraciones independientes. En este ejemplo la frase contiene una aposición explicativa. Su simplificación consiste en dos frases donde la explicación ha sido expresada en con una frase independiente.

El tercer problema consiste en simplificar el vocabulario del texto. En general las palabras poco frecuentes o que se consideran complicadas son reemplazadas por palabras sinónimas más comunes. En este caso la palabra "facultativo" es reemplazada por si sinónimo más común "médico".

Ahora bien, cómo podemos medir la complejidad o legibilidad de un texto?

Los llamados índices de legibilidad son fórmulas que asocian un valor numérico a un texto indicando su dificultad. Existen varios índices para el inglés tales como la fórmula Flesch Reading Ease o el índice Flesch-Kincaid. Estos índices combinan dos parámetros por un lado la longitud promedio de las oraciones del texto y por otro lado el número promedio de sílabas de las palabras del texto.

- "Flesch Reading Ease" (Flesch, 1949) se calcula de la siguiente manera:
    - S=206.835 – (1.015*ASL) –(84.6*ASW)
    - ASL = longitud promedio de las oraciones del texto
    - ASW = número promedio de sílabas de las palabras del texto
- Flesch-Kincaid (Kincaid et al, 1986)
    - TL = (.39 x ASL) + (11.8 x ASW) - 15.59

Sin embargo, un número de trabajos recientes en TLH consideran factores adicionales a los establecidos por los índices tradicionales. Por ejemplo, los trabajos actuales introducen

- modelos de lenguaje (es decir vocabulario),
- información sintáctica (como la profundidad del árbol sintáctico de las oraciones),
- y la información discursiva (por ejemplo que tipo de relaciones discursivas están presentes en el texto).

En general un sistema de simplificación de textos deberá incorporar alguna o todas estas tecnologías:

- Identificación de oraciones, palabras, frases, etc. pues se opera en este tipo de unidades
- Etiquetado morfosintáctico y análisis sintáctico ya que se debe determinar distintos tipos de palabras y frases.

- Reconocimiento de entidades nombradas y resolución de la correferencia dada la necesidad de reemplazar ciertas correferencias por sus antecedentes.

Los sistemas de reglas consisten de un patrón y una acción a realizar. El patrón debe ser buscado en la oración a simplificar de manera a instanciar una serie de variables. Una vez instanciado la acción indica como nuevas oraciones serán producidas.

En este ejemplo, el patrón, que contiene restricciones sintácticas y léxicas, busca una frase de relativo, que es transformada en una oración independiente. La variable que hace match con el sujeto de la frase, es utilizada para copiar el sujeto de la oración principal a la relativa.

La simplificación ha sido tratada recientemente como un problema de traducción entre un lenguaje normal y un lenguaje simple. Aquí se ven algunos ejemplos tomados de la Wikipedia en inglés y de su versión simplificada. Como estos claramente muestran, las oraciones originales y las simplificadas no son muy diferentes, aunque este fenómeno depende del recurso utilizado.

En general para aplicar un modelo de traducción automático a la simplificación de textos se necesita:

- Corpus de textos originales y simplificados.
- Alineación de oraciones y frases entre los textos originales y simplificados.
- Modelo de traducción para pasar de las oraciones complejas a las simples.
- Modelo de lenguaje simple para decidir entre múltiples soluciones.

Como ya hemos adelantado, la simplificación léxica consiste en reemplazar palabras raras por otras más comunes y que tengan el mismo sentido.

Veamos esta frase y supongamos que consideramos la palabra “hogar” como compleja, nos gustaría reemplazarla por la palabra “casa”.

- Llegó al **hogar** y se dio un baño…
    - Llegó a la **casa** y se dio un baño.

Pero ahora en este segundo ejemplo también tenemos la palabra “hogar”, sin embargo no podríamos reemplazarla por “casa” pues en este contexto no tienen el mismo sentido.

- Puso la ropa junto al **hogar** para secarla…
    - Puso la ropa junto a la **chimenea** para secarla.

Se hace necesario entonces utilizar un diccionario de sinónimos organizado por sentidos que nos indique que sinónimos existen para cada sentido de una palabra dada.

En este caso también será necesario un proceso de desambiguación de sentidos para escoger la lista de sinónimos apropiada y un criterio para escoger un reemplazo apropiado a partir de la lista de sinónimos.

El sistema LexSiS para la simplificación léxica en español usa un modelo de vectores de palabras para efectuar la desambiguación de sentidos y un criterio de simplicidad que combina la frecuencia de la palabra con la longitud.

Hay dos temas de tecnologías de lenguaje humano relacionadas con la simplificación de textos. Por un lado el resumen automático que busca reducir un texto a su contenido esencial.

Por otro lado los sistemas de extracción de definiciones que sirven para crear diccionarios que pueden ser utilizados para ofrecer clarificaciones de términos complejos o nombres a la hora de simplificar.

Eso es todo por hoy. Muchas gracias por haber escuchado esta lección.