GENERADOR INTELIGENTE DE DOCUMENTOS DE FORMACIÓN

Área temática III o IX
(Infoingeniería lingüística aplicada a la generación de documentos para la formación universitaria)

Guillermo Barrutieta Anduiza
Loramendi, 4
20500 Arrasate (Gipuzkoa)
gbarrutieta@eps.muni.es

Esta ponencia se enmarca en la generación automática e inteligente de documentos que se adapten a las necesidades de los usuarios. El contexto concreto de los documentos es el de formación para estudiantes de Universidad en un entorno trilingüe (inglés, castellano y euskara).

El prototipo que se presenta cuenta como fuente para la generación de un corpus etiquetado en XML donde las etiquetas o metadatos explicitan la estructura del discurso de acuerdo a la teoría de análisis del discurso propuesta por W. C. Mann y S.A. Thompson denominada RST.

PROBLEMÁTICA

Este proyecto se enmarca dentro del esfuerzo continuado de la comunidad científica en torno al problema de la automatización de la generación de textos 'a medida' con elementos multimedia para atender las necesidades concretas de los usuarios. Para el desarrollo de esta ambiciosa área de investigación se aplican, entre otros, métodos de Inteligencia Artificial y Generación del Lenguaje Natural enmarcados en el contexto de la Lingüística Computacional.

Este proyecto se va a centrar en la generación automática e inteligente de los documentos. Existen multitud de contextos en los que se puede detectar la necesidad de generación automática e inteligente de documentos:

- Medicina **[Hirst et al., 1997]**: Por ejemplo, un médico que escribe un informe sobre cierto paciente lo redactará de distinta manera cuando se dirija a un colega, al paciente o a un asistente técnico sanitario aunque en todos esos casos la fuente de la información pude ser la misma.

- En un contexto universitario la visión que tiene el personal administrativo de una asignatura es distinta a la que puede tener el profesor titular de la misma, un alumno que está cursando dicha asignatura u otro que está decidiendo si se matricula o no.

- De esta misma manera se podrían encontrar otros contextos en los que desde una misma fuente de información se debe extraer una vista de un documento que resulte relevante para cierto tipo de usuario en un momento dado y otras posibles vistas para otros tipos de usuario en otros casos distintos.

El contexto concreto para el que se va a desarrollar este proyecto y también un sistema sobre el que realizar la experimentación, las pruebas y validación de la teoría es el contexto relativo a documentos de formación de estudiantes de Universidad en un entorno trilingüe (inglés, castellano y euskara). De ahí que se hable en este caso de generación multilingüe.

Una de las ideas clave del sistema propuesto para dar respuesta a esta necesidad de generar documentos a medida es la utilización de un documento maestro (o *master document*, en inglés) **[Hirst et al., 1997]** multilingüe del que se extraen piezas de texto y otros elementos multimedia para generar un documento coherente y asimilable que atienda satisfactoriamente las necesidades de información de los usuarios del sistema (alumnos antes del curso, alumnos durante el curso, alumnos después del curso, profesor titular, profesor asistente, otros profesores, administración o secretaría académica, etc.) en todo momento y de acuerdo con los aspectos del perfil de un usuario dado. El documento maestro al que se hace referencia aquí se puede entender también, y utilizando términos puramente informáticos, como una base de datos documental, o en su caso, utilizando términos lingüísticos, corpus documental.

De este planteamiento surgen tres asuntos que constituyen los puntos focales del esfuerzo investigador que se va a realizar dentro del plan de trabajo de este proyecto en el área de la Lingüística Computacional:

- El modelado y creación no automáticos del documento maestro multilingüe (inglés, castellano y euskara) del que generar los textos a medida con elementos multimedia.

    - ¿De qué partes se compone un documento para formación?¿Qué dependencias y relaciones existen entre esas partes?¿Cuál es la granularidad de las partes?¿Cómo se modela el discurso de una explicación?¿Cómo se modela un ejercicio interactivo destinado a medir el grado de asimilación de ciertos conceptos?¿De qué manera afecta en el modelado del documento maestro el hecho de que se trate de un entorno multilingüe, por ejemplo, la representación es isomórfica para distintos idiomas?

Este modelado y creación del documento maestro es una parte fundamental de este proyecto y constituye un área activa de investigación **[Marcu et al., 1999]** pero no constituye la parte central en sí. Este trabajo parte de la existencia de este documento (creado manualmente) **[Marcu, 1999]** que será del que se generará el lenguaje natural o, en otras palabras, del que se generarán los documentos que contienen el lenguaje natural.

- La selección del contenido y la ordenación del discurso en función del perfil de entrada para atender a las necesidades concretas de un usuario en un momento dado.

    - ¿Cuáles son los aspectos del perfil de usuario que nos permiten hacer una selección más relevante para cada uno de ellos?¿Los aspectos del perfil de usuario son generalizables y se pueden aplicar a otros dominios diferentes al propuesto?¿Cuáles son las reglas que como humanos se aplican para seleccionar cierto discurso frente a otros también posibles en un mismo contexto?¿Cómo se modelan y utilizan estas reglas en un sistema informático?

- La selección de la presentación del texto y elementos multimedia generados para maximizar la asimilación del mismo por parte del usuario.

    - ¿Resulta más fácilmente asimilable por los usuarios un texto ordenado en frases separadas por puntos y párrafos frente a un texto ordenado en viñetas o en una tabla?¿Cuáles son las reglas que como humanos se aplican para seleccionar cierto modo de presentación de la información textual o multimedia frente a otros también posibles en un mismo contexto? ¿Cómo se modelan y utilizan estas reglas en un sistema informático?

SOLUCIÓN PREVISTA

Tal y como se ha mencionado en el apartado en el que se explicaba la problemática, este proyecto va a desarrollar su esfuerzo investigador en 3 áreas concretas:

1. Modelado y creación del documento maestro
2. Selección de contenido
3. Selección de formato

Como resultado concreto se dispone ya de un prototipo concreto que se denomina Course View Generator.

Estructura general del sistema y parámetros de entrada

El sistema dispone de un documento maestro del que se pueden extraer partes de texto para el usuario final. La selección de las piezas de texto y del modo en que estas piezas van a ser presentadas se realiza en función de ciertos parámetros o aspectos de entrada tales como:

- Momento en el tiempo que a su vez se puede subdividir en:

    - Antes del curso para decidir si se matricula o no
    - Durante el curso (día concreto) para aprender o estudiar un tema
    - Después del curso para, por ejemplo, repasar, preparar un examen, ...

- Idioma para poder elegir desde cuál de los corpus paralelos se va a generar para poder así obtener documentos en alguno de los idiomas posibles (inglés, castellano y euskara)
- Tema del que se trata
- Aspectos de usuario que a su vez se pueden subdividir en:

- Tiempo disponible
- Nivel de conocimiento previo que aumenta conforme avanza el curso
- Requemiento de trabajo participativo o ejercicios en contexto del proyecto
- Técnicas de enseñanza-aprendizaje que a su vez se puede subdividir en:

    - Repetición, repaso, esquema y resumen
    - Comparación positiva (analogía, ejemplo, ...)
    - Comparación negativa (antitesis, concesión, ...)
    - Visualización gráfica de relaciones, entidades, conceptos, ...
    - Ejercicios y experimentación
    - Razonamiento (justificación, evidencias, razones,...)
    - Contexto (antecedentes, elaboración de ideas, ...)
    - Persuasión (motivación, propósito, ...)

Esta lista de parámetros o aspectos de entrada es provisional. Lo más probable es que cambie, es decir, se ampliará, se reducirá o modificará en el transcurso del proyecto y conforme se conozca más sobre la teoría.

Básicamente el sistema va a generar distintas vistas del documento maestro, es decir, documentos para cada posible usuario, en función de la lista de parámetros detallada arriba.

Modelado del corpus multilingüe paralelo

Siempre que se trabaja en temas relacionados con la Generación del Lenguaje Natural se plantea una misma pregunta, una pregunta clásica en este tipo de proyectos de investigación: ¿De dónde generar el texto? **[Dale et al., 1998]** Es indudable que se necesita partir de una representación lingüística (o no) del lenguaje a partir de la cual generar. Nótese que, en el caso de este sistema, al tratarse de una representación lingüística se debería hablar de extracción más que de generación aunque en este informe se utilizan ambos términos de forma indistinta.

En el caso de este proyecto, se va a generar desde un corpus multilingüe paralelo que está implementado en XML (eXtensible Markup Language) **[Bray, 1998].** La organización de la información dentro del fichero XML se hace con DTDs (Document Type Definition) **[Bosak, 1997-1999]** que son parte de la tecnología XML. Y, finalmente, la representación lingüística del texto se hace siguiendo Rhetorical Structure Theory (RST) de W. Mann y S. Thompson **[1987].**

La razón de la utilización de XML-DTD es que se trata de un estándar de Internet con lo que se consigue un sistema compatible con la más avanzadas técnicas de desarrollo de contenidos web y la posibilidad de visualizar los documentos desde un navegador de páginas web de Internet. Esto facilitaría, si se llega a producir, su utilización en un entorno educativo real.

La razón de la utilización de RST es que permite representar el texto y las relaciones retóricas que se dan entre distintas partes del texto al nivel del discurso **[Mann y Thompson, 1987].** Esta teoría de descripción de la organización del texto es una teoría conocida y aceptada por la comunidad científica en el área de la lingüística computacional y uno de las alternativas más utilizadas para la generación del lenguaje natural **[Jurafsky y Martin, 2000]**.

De la combinación de estos elementos, XML-DTD y RST, se consigue crear un corpus robusto y flexible que permite generar vistas del documento maestro que es resumidamente lo que se pretende hacer con el modelado del corpus multilingüe paralelo.

Selección del contenido

La extracción técnicamente se realiza con Javascript que permite la asociación dinámica de diferentes XSLs a un mismo XML. Son los ficheros XSL **[Adler, 2000] [Clark, 1999] [Lilley et al., 1997-2000]** los que se encargan de procesar al más bajo nivel los ficheros XML para obtener las pieza de texto adecuadas en cada caso.

La estrategia y planificación del discurso se realiza utilizando las reglas de selección que se habrán modelado informáticamente y utilizando como entrada los aspectos y parámetros del sistema que nos dan a conocer el perfil del usuario. La selección de una pieza de texto frente a otra se realiza gracias al conocimiento que se tiene de la función comunicativa que cumple ese texto a nivel del discurso; este conocimiento está disponible para el sistema gracias a la utilización de Rhetorical Structure Theory (RST) **[Mann y Thompson, 1987]** que representa, precisamente, la función comunicativa que cumple cada pieza de texto y su relación con otras piezas de texto.

Selección de la presentación

La presentación se realiza con ficheros XSL **[Adler, 2000] [Clark, 1999] [Lilley et al., 1997-2000]** asociados a etiquetas HTML.

Habrá que elegir cuál es el mejor formato para mostrar un documento cuando el dispositivo de visualización es un navegador de páginas web de Internet, un PDA, un móvil con tecnología WAP, UMTS, ...

Habrá que elegir también cómo nos interesa ver las piezas de texto seleccionadas: frases separadas por puntos y agrupas en párrafos, si una pieza de texto va a ser un título o no, si el texto va a aparecer en negrita, hiperenlaces, utilización de viñetas, tablas, ...

Habrá que elegir de entre los distintos tipos de formatos aquellos que se ajusten de forma más adecuada a la función comunicativa. Por ejemplo, se ve claramente la diferencia entre un folleto promocional, un documento que recoge la información académica de un curso para su utilización por administración, un material didáctico dirigido al alumno o las notas que utiliza un profesor para impartir la clase aunque todos estos documentos pueden tener un mismo origen, el documento maestro. Este va a ser el caso en este proyecto.

Plan de validación

Este sistema se utilizará para desarrollar una serie de experimentos con los usuarios que nos van permitir validar el modelado informático de las reglas para la generación de documentos que incluye:

- Las reglas para el modelado del documento maestro
- Los parámetros de entrada que perfilan a los distintos usuarios.
- Las reglas para la selección del contenido
- Las reglas para la selección del formato

Los documentos generados van a ser valorados por los usuarios como método de validación de resultados. Estos usuarios van a evaluar la validez, es decir, la utilidad del documento en cuestión para su asimilación en el contexto de un curso de formación o de forma más general la relevancia del documento de acuerdo a sus necesidades.

De estas pruebas y sus sucesivas iteraciones se puede desarrollar una teoría que permita validar la hipótesis de partida siguiente:

- La compresión del conjunto de aspectos que sustentan las variaciones en contenido y formato de la información para diferente tipo de gente en ciertas clases de comunicación

técnico-educativa. Estos aspectos incluyen disponibilidad de tiempo, nivel de conocimiento, trabajo participativo, ...

- Las regularidades en la selección de contenido y formato de presentación pueden ser capturadas y modeladas en reglas y ejecutadas en un ordenador. Se puede por lo tanto comprobar la validez de la teoría (aspectos, reglas y material) haciendo que usuarios juzguen la relevancia del resultado presentado por el sistema.

Además en su momento se procederá a la aplicación de esta teoría en otro dominio diferente al utilizado durante este proyecto para verificar la generalidad de la teoría.

BIBLIOGRAFÍA

Adler, S [2000] Extensible Stylesheet Language (XSL) Version 1.0. *World Wide Web Consortium* <URL: http://www.w3.org/TR/xsl/>

Bosak, J. et al. [1997-1999] W3C XML Specification DTD ("XMLspec") *World Wide Web Consortium* <URL: http://www.w3.org/XML/1998/06/xmlspec-report.htm>

Bray, T. et al. (eds) [1998] Extensible Markup Language (XML) 1.0. *World Wide Web Consortium* <URL: http://www.w3.org/TR/REC-xml>

Clark, J. (ed) [1999] XSL Transformations (XSLT) Version 1.0 *World Wide Web Consortium* <URL: http://www.w3.org/TR/xslt>

Dale, R., DiEugenio, B., Scott, D. [1998] Introduction to the Special Issue on Natural Language Generation. *Association for Computational Linguistics. MIT Press*. Cambridge, MA.

Hirst, G., DiMarco, C., Hovy E., Parsons K. [1997] Authoring and Generating Heath-Educacion Documents That Are Tailored to the Needs of the Individual Patient. *Proceedings of the Sixth International Conference, UM97*. Vienna, NY.

Jurafsky, D., Martin, J.H. [2000] Speech and Language Processing. *Prentice Hall.* Upper Saddle River, NJ.

Lilley, C., Quint, V. (eds) [1997-2000] Extensible Stylesheet Language (XSL) *World Wide Web Consortium* <URL: http://www.w3.org/Style/XSL/>

Mann, W.C., Thompson, S.A. [1987] Rhetorical Structure Theory: A theory of text organization. *Tech. Rep. RS-87-190. Information Sciences Institute*. Los Angeles, CA.

Marcu, D., Romera, M., Amorrortu, E. [1999]. Experiments in Constructing a Corpus of Discourse Trees: Problems, Annotation Choices, Issues. *The Workshop on Levels of Representation in Discourse, pages 71-78.* Edinburgh, Scotland.

Marcu, D. [1999] Instructions for Manually Annotating the Discourse Structures of Texts. *ISI-USC*. Los Angeles, CA.