Sistemas Interativos Multimodales de Procesamiento del Lenguaje Natural *

Natural Language Processing Interactive Multimodal Systems

Elsa Cubel, Alejandro H. Toselli
Instituto Tecnológico de Informática
Universidad Politécnica de Valencia
Camino de Vera s/n, 46022, Valencia
{ecubel, ahector}@iti.upv.es

Resumen: En este trabajo se plantea una aproximación novedosa en la que los sistemas de PLN cooperan conjuntamente con el usuario en el procesamiento y consecución satisfactoria de la tarea.
Palabras clave: interacción multimodal, transcripción, traducción automática, recuperación de imágenes

Abstract: In this work, a novel approach is introduced in which NLP systems cooperate together with users in the processing and satisfactory achievement of a given task.
Keywords: multimodal interaction, transcription task, machine translation, image retrieval

Resumen: En los últimos años ha tenido lugar un gran avance, tanto en el desarrollo de tecnologías multimodales interactivas como en el de interfaces avanzadas persona-máquina en el campo del *procesamiento de lenguaje natural* (PLN). Especialmente, las áreas del *reconocimiento de formas* y *visión por computador* vienen jugando un papel preponderante en el desarrollo de este tipo de tecnologías e interfaces.

Actualmente, se considera que la total automatización que presentan los sistemas tradicionales de PLN, no resulta lo más conveniente cuando se requieren resultados completamente libres de errores. Por el contrario, en este trabajo se plantea una aproximación novedosa en la que los sistemas de PLN cooperan conjuntamente con el usuario en el procesamiento y consecución satisfactoria de la tarea.

Como ejemplos de esta aproximación novedosa, se describen algunas aplicaciones muy usuales en PLN, como son la transcripción de textos manuscritos y señal de audio, traducción automática y recuperación de contenidos multimedia.

* Trabajo financiado parcialmente por la EC (FEDER/FSE) y el MEC/MICINN español en el marco del proyecto MIPRCV (CSD2007-00018) bajo el programa "Consolider Ingenio 2010", los proyectos iTrans2 (TIN2009-14511) y MITTRAL (TIN2009-14633-C03-01).

1. Introducción

El *Procesamiento del Lenguaje Natural* (PLN) se ocupa de proveer métodos y técnicas que automáticamente faciliten la comunicación entre personas o entre personas y máquinas por medio de lenguajes naturales. Entre las líneas de investigación o aplicaciones atribuidas al PLN, podemos citar el de la síntesis del habla, reconocimiento de voz, traducción automática, reconocimiento de texto manuscrito, etc.

Tradicionalmente, en estas líneas de investigación, los métodos y técnicas de PLN utilizados se centraban en el desarrollo de aplicaciones totalmente automatizadas. Sin embargo, dado que los resultados de la mayoría de las mismas distan mucho de ser perfectos, una intervención humana experta (que denominaremos `usuario` de ahora en adelante) era finalmente requerida para la validación de los mismos. En este caso, los usuarios suelen utilizar las aplicaciones de PLN de este tipo, dentro de un proceso de dos etapas: en primer lugar, la aplicación procesa automáticamente toda la tarea; y a continuación, el usuario revisa y corrige sus resultados para que la calidad final sea aceptable. Este proceso es lo que se conoce como *post-edición*. Este proceso, aunque permite obtener resultados de

calidad, resulta por lo general bastante ineficiente e incómodo para el usuario, quien podría preferir prescindir de la salida de la aplicación y procesar la tarea directamente desde cero y por sí mismo.

Como alternativa, se propone un enfoque más pragmático, conocido como *paradigma interactivo-predictivo* (IP), en el cual tanto la aplicación de PLN como el usuario colaboran mutuamente para completar la tarea de manera eficiente. De este modo, se consigue combinar en un mismo sistema, la eficacia (en términos de rapidez) de las aplicaciones de PLN tradicionales, con la precisión aportada por la experiencia del usuario. En este sentido, en la última década la demanda social e industrial de tecnologías interactivas multimodales para el desarrollo de interfaces avanzadas hombre-máquina ha crecido considerablemente. Especialmente, las áreas del *reconocimiento de formas* y *visión por computador* han venido jugando un papel preponderante en el desarrollo de este tipo de tecnologías e interfaces.

En este trabajo presentamos varias tecnologías IP-PLN, implementadas en diferentes prototipos completamente funcionales de aplicaciones, que muestran *in situ* los beneficios de cada una de ellas. El desarrollo e implementación de estos prototipos se ha focalizado también en el paradigma de la *multimodalidad*, posibilitando que el usuario pueda interactuar de forma más natural y ergonómica con dichos prototipos.

2. Paradigma Interactivo-Predictivo Multimodal

En el marco del proyecto nacional "Multimodal Interaction in Pattern Recognition and Computer Vision" (MIPRCV Consolider-Ingenio 2010), se vienen desarrollando tecnologías bajo el nuevo paradigma IP multimodal del que hemos hablado. Todos los prototipos desarrollados en este proyecto están basados en estas tecnologías y para la mayoría de ellos (principalmente los relacionados con PLN) se ha establecido una forma de interacción común del usuario con los mismos. El objetivo es poder emplear un mismo protocolo de interacción con estos prototipos de aplicación, disminuyendo así la carga cognitiva del usuario y facilitando un rápido aprendizaje en la utilización del sistema. Básicamente, este protocolo establece el modo en que se va a llevar a cabo la interacción aplicación-usuario conforme se va procesando una determinada tarea. En otras palabras, a medida que la aplicación va mostrando resultados parciales, el usuario podrá (mediante acciones) proceder a su validación, corrección, etc.; y, posteriormente, la aplicación, en base a estas acciones del usuario, podrá ofrecer nuevos resultados alternativos.

Las bases sobre las que se ha fundamentado la implementación de los prototipos son las siguientes:

i- Realimentación del usuario: Las acciones correctivas propuestas progresivamente por el usuario con cada propuesta de resultados, son realimentadas al sistema introduciendo restricciones de contexto que ayudan a sugerir nuevas propuestas de resultados más precisas.

ii- Aprendizaje adaptativo: Se aprovechan las acciones correctivas introducidas por el usuario para adaptar progresivamente *in situ* los modelos de la tarea, que serán utilizados por la aplicación para proponer mejores resultados.

iii- Multimodalidad: La multimodalidad aparece en estos sistemas de forma natural. Las acciones del usuario destinadas a corregir los resultados que son presentados por la aplicación en cada momento, pueden provenir de múltiples modos: desde las tradicionales pulsaciones de teclado o movimientos del ratón a sistemas de reconocimiento del habla o de gestos.

3. Demostradores

En esta sección se describen algunos de los prototipos de aplicaciones basados en tecnologías IP-PLN multimodales desarrollados en el marco del proyecto nacional MIPRCV Consolider-Ingenio 2010[1].Como se observará, todos estos sistemas funcionan siguiendo el paradigma interactivo-predictivo multimodal, el cual introduce totalmente al usuario como una parte más del sistema.

Los prototipos comparten una arquitectura cliente-servidor sobre Internet (Alabau et al., 2009).

[1] `http://miprcv.iti.upv.es`

Figura 1: Interfaces de prototipos MM-CATTI (izquierda) and CAST (derecha).

3.1. Prototipos de Transcripción y Traducción Interactiva Multimodal

En esta sección presentamos dos prototipos de transcripción, completamente funcionales, destinados a la transcripción de imágenes de texto manuscrito (Toselli et al., 2009) y señal de audio (Rodríguez, Casacuberta, y Vidal, 2007) respectivamente. También presentamos un prototipo destinado a la traducción de textos (Casacuberta et al., 2009). Todos estos prototipos se han desarrollado e implementado siguiendo el paradigma IP multimodal que hemos presentado previamente. En estos prototipos, el usuario interactúa con el sistema validando segmentos correctos de transcripción/traducción y corrigiendo sus subsiguientes errores. A continuación, teniendo en cuenta estos segmentos validados y las correciones efectuadas, el prototipo genera mejores sugerencias de transcripción/traducción en la siguiente interacción. El usuario puede realizar las mencionadas validaciones y correcciones mediante el teclado y ratón, o por medio de otras modalidades de interacción más sofisticadas como lápiz electrónico (escritura *on-line*) o reconocimiento del voz.

El prototipo de transcripción de imágenes de texto manuscrito, denomindo "Multimodal Computer Assisted Transcription of Text Images" (MM-CATTI) (figura 1 - izquierda), se encuentra accesible en: http://catti.iti.upv.es. A través del mismo, se podrá experimentar con la transcripción interactiva multimodal de documentos de diferente naturaleza: documentos manuscritos antiguos (*Cristo Salvador* del siglo XIX), texto manuscrito moderno (IAMDB en inglés), escritura manuscrita realizada en formularios de encuestas, etc.

Por su parte, el prototipo de transcripción de señal de audio, denominado "Computer Assisted Speech Transcription" (CAST) (ver figura 1 - centro), resulta de gran interés en diversas aplicaciones como: subtitulado de programas de televisión, accesibilidad a personas con discapacidad auditiva, búsquedas textuales de contenidos de audio, transcripciones de programas de radio, conferencias, sesiones judiciales, etc.

Para ambos prototipos, MM-CATTI y CAST, de acuerdo a los resultados experimentales, cuando se compara el sistema de transcripción basada en el paradigma IP multimodal con una transcripción manual completa, la reducción estimada de esfuerzo del usuario está entre un 68 % y un 80 %.

Por otro lado, el prototipo web para la traducción interactiva (ver figura 2 - izquierda), está disponible en: http://cat.iti.upv.es/imt. Según los experimentos llevados a cabo con este prototipo, el usuario reduciría hasta en un 30 % el esfuerzo necesario hasta alcanzar la traducción correcta si lo comparamos a la utilización de un sistema totalmente automático. Las aplicaciones que puede tener este prototipo son múltiples: traducción de manuales, traducción de textos oficiales, traducción de páginas web, etc.

3.2. Prototipo de Recuperación Interactiva de Contenidos Multimedia

En las consultas de colecciones con contenidos multimedia, utilizando sistemas convencionales de recuperación de información, se buscan aquellos contenidos que más se asemejan a la consulta realizada. Muchas veces la información recuperada con estos sistemas no cubre las expectativas del usuario; en parte debido a la propia falta de información específica de la consulta realizada. Sin embargo, si se utiliza el paradigma IP

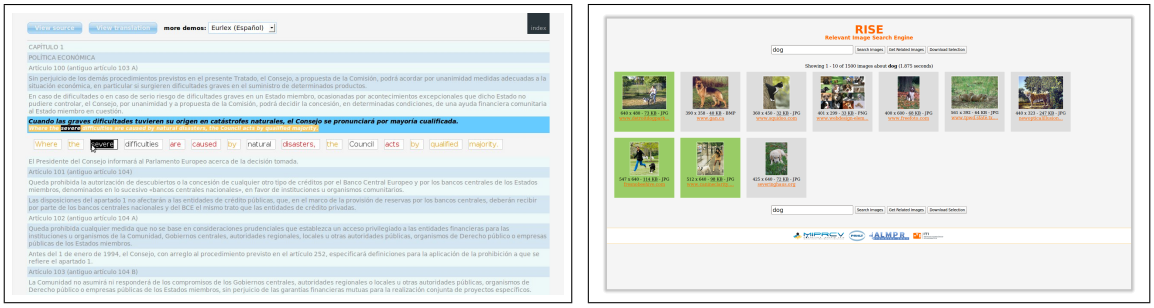

Figura 2: Interfaces de prototipos CAT (izquierda) and RISE (derecha).

multimodal, el usuario puede proporcionar una retroalimentación relevante sobre la adecuación de la información recuperada.

En http://rise.iti.upv.es puede experimentarse con el prototipo web de recuperación interactiva de contenidos multimedia (figura 2 - derecha), denominado "Relevant Image Search Engine" (RISE) (Cevikalp y Paredes, 2009). Este prototipo de aplicación es un buscador de imágenes donde, en primer lugar, el usuario introduce el término que desea buscar. La aplicación trabaja como un interfaz con Google Images, que es quien provee las imágenes a partir de los términos de la búsqueda que ha introducido el usuario. El usuario selecciona aquellas imágenes que considera que más se ajustan a lo que desea ver y a partir de entonces, iterativamente, el sistema devolverá aquellas imágenes que sean más relevantes a partir de la selección del usuario. A modo de ejemplo, en la figura 2 - derecha sabemos que el usuario pretende encontrar imágenes de perros que llevan collar. Cada vez que la aplicación muestra una respuesta, el usuario solamente seleccionará aquellas imágenes en las que aparezcan perros con collar (las tres imágenes que aparecen seleccionadas en la figura). De esta forma, en pocas interacciones se conseguirá que la aplicación solamente muestre imágenes que cumplan los requisitos del usuario.

4. Conclusiones

En este trabajo se ha presentado el paradigma interactivo-predictivo multimodal bajo el cual, un sistema de PLN facilita y colabora conjuntamente con el usuario en la producción de resultados de alta calidad. En este contexto, se han presentado diversos prototipos, completamente funcionales, que ejemplifican áreas de aplicación de gran interés e importancia: transcripción de imágenes de texto manuscrito y señal de audio, traducción de textos y recuperación de contenidos multimedia.

En todos los casos, se ha constatado que los prototipos diseñados bajo este nuevo paradigma, reducen significativemente el esfuerzo que el usuario debe realizar para alcanzar un resultado correcto.

Bibliografía

[Alabau et al.2009] Alabau, V., D. Ortiz, V. Romero, y J. Ocampo. 2009. A multimodal predictive-interactive application for computer assisted transcription and translation. En *ICMI-MLMI '09: Proceedings of the 2009 international conference on Multimodal interfaces*, páginas 227–228, New York, NY, USA. ACM.

[Casacuberta et al.2009] Casacuberta, F., J. Civera, E. Cubel, A.L. Lagarda, G. Lapalme, E. Macklovitch, y E. Vidal. 2009. Human interaction for high quality machine translation. *Communications of the ACM*, 52(10):135–138.

[Cevikalp y Paredes2009] Cevikalp, Hakan y Roberto Paredes. 2009. Semi-supervised distance metric learning for visual object classification. En *VISSAPP (1)*, páginas 315–322.

[Rodríguez, Casacuberta, y Vidal2007] Rodríguez, L., F. Casacuberta, y E. Vidal. 2007. Computer Assisted Transcription of Speech. En *Proceedings of the 3rd Iberian Conference on Pattern Recognition and Image Analysis*, volumen 4477 de *LNCS*, páginas 241–248, Girona (Spain), June.

[Toselli et al.2009] Toselli, Alejandro H., Verónica Romero, Moisés Pastor, y Enrique Vidal. 2009. Multimodal interactive transcription of text images. *Pattern Recognition*, 43(5):1814–1825.