

CHISPA-MT: un Consorcio para el Desarrollo de Sistemas de Traducción Automática entre las Lenguas China y Castellana

Rafael E. Banchs

Universitat Politècnica de Catalunya
Jordi Girona 1-3, D4-100, Campus Nord
UPC, 08034 Barcelona, Spain
rbanchs@gps.tsc.upc.edu

Resumen: El presente proyecto de investigación pretende el estudio y desarrollo de tecnologías para la traducción automática en sistemas de comunicación chino-castellano y castellano-chino. El proyecto combina una línea de investigación básica con una línea de investigación aplicada cuyo objetivo es la utilización de las tecnologías generadas en el desarrollo de herramientas específicas para la comunicación bilingüe.

Palabras clave: Traducción Automática Estadística, Comunicación Bilingüe, Chino, Castellano

Abstract: The main goal of this research project is the study and development of machine translation technologies for Chinese-to-Spanish and Spanish-to-Chinese communication systems. The project combines a basic research activity along with an applied research work which objective is the implementation of the developed machine translation technologies into specific tools and systems for bilingual communications.

Keywords: Statistical Machine Translation, Bilingual Communication, Chinese, Spanish

1 Motivación

No obstante todo el esfuerzo dedicado actualmente a la investigación y desarrollo de sistemas de traducción automática, se puede constatar que la mayor cantidad de dicho esfuerzo está concentrando en pares de lenguas tales como: chino-inglés, chino-francés, árabe-inglés, castellano-inglés, etc. (Gale,2005; Manos,2004; Tc-star,2004), mientras el esfuerzo en pares de lenguas como el chino-castellano es casi nulo.

En este sentido, resulta contradictorio el casi inexistente esfuerzo dedicado en los actuales momentos a la investigación y desarrollo de sistemas de traducción automática entre chino y castellano, y el creciente número de acuerdos y convenios, tanto comerciales como científico-tecnológicos, que en los últimos años han estado suscribiendo España y muchos países latinoamericanos con la República Popular China. De esta forma, la investigación y desarrollo de sistemas de traducción para este par específico de lenguas, más que de un interés científico en sí mismo, es de gran interés político y económico para todos los países de habla hispana.

En este contexto, el proyecto CHISPA-MT se propone cubrir esta necesidad existente en investigación orientada a la traducción automática entre el par específico de lenguas china y castellana, así como el desarrollo de tecnologías específicas de comunicación bilingüe chino-castellano.

2 Características del Consorcio

El proyecto CHISPA-MT¹ está concebido bajo la figura de un consorcio académico de financiación privada, cuyo financiamiento se realiza a través de la participación de diversas empresas e instituciones patrocinantes que se benefician tanto de la comercialización como de la obtención de licencias para todas las aplicaciones y tecnologías que se desarrollan en el consorcio. Tres distintos tipos de socios participan en el consorcio:

¹ Cabe destacar que el proyecto está en estos momentos en su fase más inicial de desarrollo, y aunque sus líneas de investigación principales se encuentran ya activas, todavía quedan por definirse algunos de los grupos participantes.

1.- Socios investigadores: son universidades y/o centros de investigación que se asocian al programa para participar en ciertas actividades de investigación o desarrollo.

2.- Socios explotadores de la tecnología: son empresas o instituciones que están interesadas en obtener beneficios económicos a través de la distribución y comercialización de servicios y productos relacionados con el uso de las aplicaciones desarrolladas por el consorcio.

3.- Socios usuarios de la tecnología: son instituciones y/o empresas que están interesadas en beneficiarse mediante el uso de las distintas aplicaciones desarrolladas por el consorcio en sus actividades internas.

3 Líneas de Investigación

En cuanto a líneas de investigación, el proyecto comprende una primera línea de investigación básica cuyos objetivos fundamentales son la mejora del estado actual del arte en sistemas de traducción automática estadística, así como la adecuación de los sistemas existentes a los problemas específicos relacionados con las tareas de traducción chino-a-castellano y castellano-a-chino.

Adicionalmente, el proyecto también contempla una segunda línea de investigación aplicada cuyo objetivo principal es la utilización de las tecnologías generadas por la línea de investigación básica para el desarrollo de herramientas específicas de comunicación bilingüe entre chino y castellano: aplicaciones para traducción humana asistida, herramientas para navegación bilingüe en Internet, plataformas de traducción para mensajería electrónica (SMS, e-mail, chat), y aplicaciones de apoyo para vídeo-conferencia bilingüe.

Cabe destacar que, para esta línea de investigación aplicada, es de fundamental importancia la participación de la tercera categoría de socios mencionada anteriormente, ya que como usuarios finales de la tecnología ellos permitirán la realización de las correspondientes pruebas de *usabilidad* de las herramientas desarrolladas en el proyecto. Es precisamente este tipo de interacción la que garantizará una rápida y eficiente adecuación de todas las aplicaciones desarrolladas en el consorcio a las necesidades reales de los usuarios finales de la tecnología.

4 Cronograma de Actividades

El proyecto CHISPA-MT cuenta actualmente con una planificación de actividades a 5 años, la cual se presenta a continuación en una forma bastante resumida. En la lista de actividades presentada “R” se refiere a las actividades de investigación básica y “A” a las relacionadas con el desarrollo de aplicaciones y productos.

Primer año:

(R) Desarrollo de un corpus bilingüe y de un sistema de traducción base.

(A) Desarrollo de prototipos para traducción asistida y navegación bilingüe.

Segundo año:

(R) Depuración del corpus bilingüe y desarrollo de recursos lingüísticos generales.

(A) Versiones actualizadas de las herramientas de traducción y navegación, y prototipo de la plataforma para mensajería electrónica.

Tercer año:

(R) Desarrollo de recursos lingüísticos para traducción en dominios específicos.

(A) Versiones comerciales de las herramientas de traducción y navegación, versión actualizada y primeros servicios de mensajería electrónica.

Cuarto año:

(R) Desarrollo de recursos para reconocimiento y síntesis del habla.

(A) Versión actualizada y extensión completa de los servicios de mensajería electrónica.

Quinto año:

(R) Pruebas de *usabilidad* de las herramientas.

(A) Versión comercial de la plataforma de mensajería electrónica, y primer prototipo de aplicaciones de para vídeo-conferencia bilingüe.

Bibliografía

Gale, 2005, Proyecto GALE, *Global Autonomous Language Environments*, <http://ciir.cs.umass.edu/research/nightingale.html>

Manos, 2004, Proyecto MANOS, *Multilingual Application Network for Olympic Services*, <http://nlpr-web.ia.ac.cn/english/cip/english/project.htm>

Tc-star, 2004, Proyecto TC-STAR, *Technology and Corpora for Speech-to-Speech Translation*, <http://www.tc-star.org/>