

Perspectivas de un Centro Nacional de Conocimiento, Información y Tecnologías del Lenguaje

Gerardo E SIERRA
Instituto de Ingeniería
Universidad Nacional Autónoma de México
México, D.F., 04510, México

RESUMEN

Con el fin de contribuir a la generación de conocimiento que se puede obtener a partir del análisis y tratamiento de la información almacenada en textos orales y escritos, se propone en México un proyecto a nivel nacional para coordinar diversos esfuerzos que existen en investigación y desarrollo para que converjan hacia un fin común, mediante un plan de desarrollo estratégico conforme a las necesidades del país. Se presentan las necesidades que motivan este proyecto, así como los objetivos específicos y el planteamiento del Centro Nacional de Conocimiento, Información y Tecnologías del Lenguaje.

Palabras clave. Gerencia de sistemas de información, gestión de conocimiento, manejo de información, tecnologías del lenguaje.

1. INTRODUCCIÓN

La información y el conocimiento son componentes esenciales para el progreso económico y para el bienestar personal. A fin de lograr la generación de conocimiento para tener las respuestas más adecuadas ante diversas circunstancias, es necesario disponer de información oportuna y verídica.

Mediante una política inteligente de acceso a la información se logra un avance importante para una interacción eficiente entre los sectores industriales y estatales, los sectores académicos y la sociedad. Al establecer un sistema de procesamiento eficiente de intercambio de conocimiento entre las partes mencionadas, se pueden realizar acciones concretas con resultados óptimos que ayuden a los diferentes estratos de la población, por ejemplo pequeñas y medianas empresas (PyMEs) con objetivos de comercio internacional que coadyuven al desarrollo económico y sustentable del país.

Las empresas requieren tratar con la gestión de información y, a mediana o gran escala, actualizar su información y gestionarla de manera que puedan tener una mejor competencia en el mercado y generar un mayor número de divisas [1]. Igualmente, los institutos y entidades requieren de una gestión y una explotación de calidad de información eficiente, por lo que es menester tomarlos en cuenta para que México siga en su ruta de avance y progreso.

Hoy en día nos enfrentamos a una masificación de información, problema que atañe y tiene repercusiones en todos los estratos de la sociedad. Tan solo como ejemplo tómanse los inmensos volúmenes de información que inundan los acervos de Medline en medicina o de ETDEWEB para energía. Similarmente crece de manera exponencial la enorme cantidad de registros administrativos y documentos en el ámbito empresarial o gubernamental para áreas estratégicas. Toda esta información

representa un problema cada vez menos asible, lo que a la larga tendrá consecuencias importantes en cuanto al control que el Estado tenga o no de toda esta información.

El análisis de información para generar conocimiento es un proceso que involucra conocer a fondo el lenguaje natural, que es el medio por excelencia para transmisión de la misma [2]. Para ello, diversas disciplinas se han orientado al análisis de lenguaje para la obtención de información y generación de conocimiento. Entre ellas, la lingüística descriptiva y aplicada, la terminología, la semiótica, la bibliotecología y documentación, la ciencia e ingeniería en computación, el procesamiento de lenguaje natural, la inteligencia artificial y otras más que en su conjunto constituyen las tecnologías del lenguaje, de la información y del conocimiento.

2. EL MANEJO DE INFORMACIÓN, PRIORIDAD NACIONAL

Dentro de la política nacional, una de las prioridades que se han manifestado en el *Programa Especial de Ciencia y Tecnología 2001-2006* [3] es el manejo de información científica y tecnológica. Hasta ahora, existen algunas empresas nacionales que proveen la digitalización de documentos y realizan sistemas de recuperación de dichos documentos, todo con bases computacionales sólidas, pero sin entrar en la complejidad que reviste manejar los textos en lenguaje natural, utilizando las tecnologías adecuadas. Por otro lado, en México existen grupos académicos que han contribuido a crear tecnología, infraestructura, recursos y desarrollos para el manejo y explotación de información, aunque principalmente lo han hecho de forma aislada, con temas particulares y sin un fin común explícito.

Asimismo, resulta vital la creación de una plataforma para el procesamiento del lenguaje natural en español, así como el desarrollo de recursos lingüísticos en nuestro idioma, todo con miras al desarrollo de aplicaciones como sistemas de resumen automático, interfaces inteligentes, sistemas pregunta-respuesta, sistemas de extracción de información, buscadores inteligentes, entre otros que actualmente se empiezan a crear en el país.

A nivel estratégico, hace falta un proyecto a nivel nacional que rija, gestione y coordine el trabajo de investigación y desarrollo de los grupos, laboratorios y centros relacionados con las tecnologías de la información, lo que provoca una enorme desorganización e inhibe avances coordinados en la solución de problemas específicos. Como consecuencia, se tiene un estancamiento tecnológico y un elevado gasto para el país, pues se ve obligado a adquirir tecnología extranjera ante la poca o nula producción nacional, esto sin mencionar los altos costos de implementación y difusión de tecnologías de información que de inicio son ajenas a la cultura e idiosincrasia de los usuarios del país.

Finalmente, pero no menos importante, hay una falta de formación y de visión interdisciplinaria a nivel académico, a la vez que no hay promoción laboral para los profesionistas egresados tanto del nivel medio-superior, superior y posgrado. En el caso de nuestro país, la relevancia estratégica que tienen todas estas áreas es aún emergente, debido en gran medida, a la falta de conocimiento y difusión. Por ello, para lograr un alto grado de calidad y competitividad en la generación y aplicación de tecnologías de la información en el país, es prioritario fomentar y destinar fondos para el desarrollo de estas interdisciplinas en nuestro medio.

3. LAS TECNOLOGÍAS DEL LENGUAJE EN EL MANEJO DE INFORMACIÓN

La clave para el manejo futuro de la información y el conocimiento son los textos. Una prueba de ello es el impacto que tiene Internet en el trabajo cotidiano a todos los niveles, pues gestiona el acceso al conocimiento, aunque todavía de manera rudimentaria.

El tratamiento de información oral y escrita requiere de estudios teóricos del lenguaje natural (realizado por lingüistas, bibliotecólogos, comunicólogos y otros pertenecientes al área humanística) así como de una serie de metodologías y técnicas de análisis (todas, con un perfil de ciencias exactas, como computación, informática, estadística, ingeniería, etc.).

El fruto de esta fusión de disciplinas ya se ha visto concretado en el importante auge que ha tenido en otros países lo que se conoce como *tecnologías e industrias del lenguaje y la información*. Este sector ha dado lugar a una gran cantidad de áreas de investigación y estudio fundadas desde mediados del siglo pasado. Estas áreas se han consolidado hasta lograr un crecimiento exponencial en los 80 y 90. De entre ellas, figuran entre las más relevantes la inteligencia artificial, la robótica, el aprendizaje-automático, el procesamiento de lenguaje natural y de voz, la ingeniería lingüística, la lingüística computacional, la extracción y recuperación de información, la representación y generación de conocimiento, la terminología y la minería de textos.

Las tecnologías del lenguaje son todas aquellas que se integran en aplicaciones informáticas para permitir el tratamiento de textos escritos o el procesamiento del habla. Se trata, en conjunto, de tecnologías que hacen posible la creación de herramientas pensadas para auxiliarnos en el uso de las computadoras sin renunciar por ello al uso habitual del lenguaje como medio de interacción y de intercambio de información.

Estas tecnologías abren un nuevo capítulo en la historia de la sociedad de la información; pueden enriquecer nuestro entorno informático con muchas de las funciones expresivas naturales propias del lenguaje humano, que hasta ahora no habían estado presentes plenamente en la interacción entre personas y máquinas. Combinando el potencial tecnológico con la profundidad y la flexibilidad del procesamiento del lenguaje humano, las aplicaciones y los servicios que ofrecen las tecnologías de la lengua pueden mejorar radicalmente la eficacia y la facilidad de los mecanismos y sistemas usados en las actividades cotidianas relacionadas con la información y la comunicación. Esto incluye el ocio y los negocios, la administración y la educación, los servicios y la vida diaria.

Algunas aplicaciones típicas de las tecnologías del lenguaje en la actualidad son:

Servicios de traducción automática, a los que se puede acceder a través de una página web o del teléfono móvil.

Sistemas de dictado automático, que permiten transcribir cartas, informes o notas en diversos campos profesionales.

Buscadores electrónicos con la integración de recursos lingüísticos, capaces de localizar y suministrar de forma inmediata información actualizada, a partir de bases documentales multilingües.

Pantallas hablantes, que permiten a usuarios con discapacidades visuales superar las limitaciones que habían tenido hasta ahora.

Diccionarios electrónicos multilingües, que facilitan la lectura de textos en lenguas extranjeras.

4. NECESIDADES ACTUALES EN MÉXICO

Son muchas las necesidades que impulsan el desarrollo de las tecnologías del lenguaje, la información y el conocimiento, siendo tres las más importantes:

- 1) Salir del estancamiento tecnológico y disminuir el elevado gasto para el país por la adquisición de tecnología extranjera ante la poca o nula producción nacional, o ante los altos costos de implementación y difusión de tecnologías de información que de inicio son ajenas a la cultura e idiosincrasia de los usuarios del país.
- 2) Permear con la interdisciplina de la tecnología del lenguaje a las nuevas generaciones de profesionales y graduados en las áreas afines a computación y lingüística, a la vez de promover fuentes de trabajo para profesionistas egresados tanto del nivel medio-superior, superior y posgrado.
- 3) Implementar y apoyar un proyecto a nivel nacional que rijan, gestione y coordine el trabajo de investigación y desarrollo de los grupos, laboratorios y centros relacionados con las tecnologías de la información para la solución de problemas específicos.

5. EL CNCITL, UNA PROPUESTA INTEGRAL

Dada la situación actual descrita, resulta necesario concretar un proyecto a nivel nacional para contribuir a la gestión de información que conduzca al descubrimiento, generación y administración de conocimiento, con base en el uso de tecnologías de lenguaje, partiendo de la inmensa base de información existente en medios electrónicos tanto en formatos textuales como orales.

Una propuesta para ello es crear un *Centro Nacional de Conocimiento, Información y Tecnologías del Lenguaje (CNCITL)*, como centro de ámbito nacional para la investigación aplicada y desarrollo tecnológico, con carácter formativo y de vinculación empresarial, que encamine los esfuerzos para la creación de nuevas tecnologías para la gestión

de información y conocimiento con base en las tecnologías del lenguaje. El fin último será coordinar los diversos esfuerzos que existen en I+D en el área para que converjan hacia un fin común, mediante un plan de desarrollo estratégico conforme a las necesidades del país. Entre los objetivos particulares de esta propuesta se proponen:

- 1) Desarrollar productos y servicios tecnológicos para la gestión y explotación de información de diversa naturaleza; especialmente, para el descubrimiento y la generación de nuevo conocimiento con el fin de alcanzar niveles competitivos y de calidad, logrando con ello un impacto a nivel internacional.
- 2) Establecer una red multidisciplinaria de investigación básica y aplicada con los centros de investigación nacionales y su contraparte internacionales, con el fin de establecer mecanismos de transferencia tecnológica que sustenten la continua innovación de tecnologías desarrolladas.
- 3) Encauzar la investigación básica y aplicada para crear productos capaces de procesar y manejar los volúmenes gigantes de información que se requieran para generar conocimiento.
- 4) Fomentar la formación de profesionales competitivos y de recursos humanos de alta especialización en el país, sobre todo considerando que pueden llegar a ser más competitivos y rentables que sus similares de países del primer mundo.
- 5) Crear y reforzar programas de licenciatura, posgrado y actualización continua que apoyen la especialización y la competitividad de los integrantes del proyecto y de los cuadros que a largo plazo trabajen en este rubro para el gobierno y las empresas.
- 6) Facilitar los medios para captar y usar de un modo inteligente la información oral o escrita, proveniente de cualquier campo del conocimiento y de cualquier sector, sea público, privado o social.
- 7) Promover y facilitar la especialización de empresas dentro de la incipiente industria de tecnologías de la información nacional, que impulsen la adquisición y uso de dichas tecnologías a nivel nacional, entre los diferentes actores de nuestra sociedad del conocimiento.
- 8) Vincular los sectores públicos, privados y sociales para el fomento de tecnologías, recursos y productos desarrollados en el país. Asimismo, acercar la academia y la industria para la colaboración entre los dos rubros y para la optimización de la tecnología desarrollada.
- 9) Generar ingresos mediante la venta en el extranjero de la tecnología desarrollada, en lugar de causar gastos en la compra de la misma.

6. ÁREAS DEL CNCITL

El concepto del CNCITL se guía por un conjunto de metas claramente delimitadas, articulándolo en torno a tres áreas.

Investigación y desarrollo

Esta área tiene por objeto mantener líneas permanentes de investigación aplicada, principalmente en las tecnologías del lenguaje, para la mejora continua de los productos y servicios del CNCITL, así como para la búsqueda de la innovación en cada uno de dichos productos y servicios. El desarrollo tecnológico tiene por objeto generar los prototipos necesarios para dar paso a la implantación de éstos en productos existentes, o bien para la generación de nuevos productos y servicios.

Para conseguir la innovación y el desarrollo de productos y servicios, el área de investigación y desarrollo realizará actividades de investigación aplicada en las siguientes líneas específicas.

Recursos lingüísticos: La investigación y desarrollo de recursos lingüísticos permitirán enriquecer el inventario actual de recursos existentes entre los diferentes grupos de investigación a nivel nacional y contar así con mayor cantidad y mejores colecciones de léxicos, reglas, corpora, y colecciones de prueba para diferentes disciplinas, por mencionar algunos. Estos recursos son de gran importancia debido a que conforman una plataforma de soporte para los métodos de procesamiento del lenguaje natural, así como para las aplicaciones de gestión de información.

Las actividades asociadas con este rubro se enfocan en la creación, obtención y mejoramiento de los recursos tales como: Recursos léxicos (tesauros, ontologías, terminologías y diccionarios), corpus lingüísticos (textuales y orales) y herramientas de análisis lingüístico (gramáticas, lematizadores, analizadores sintácticos).

Procesamiento del lenguaje natural (PLN): La investigación y desarrollo en materia de métodos para el procesamiento del lenguaje natural permitirán el desarrollo de herramientas computacionales para el análisis y tratamiento de fuentes de información orales y escritas mediante la aplicación de dichos métodos. Esta área es de particular importancia ya que actualmente no se cuenta con una base robusta de herramientas de diseño y producción nacional. En la mayoría de los casos se depende de herramientas extranjeras, principalmente se trata de herramientas de origen académico desarrolladas por grupos de investigación españoles, razón por la cual puede decirse que son adecuadas para actividades académicas, no así para actividades de desarrollo orientadas a la creación de herramientas productivas en ambientes reales de trabajo.

Las actividades asociadas dentro del CNCITL en este rubro incluyen la creación, adecuación, aplicación y evaluación de herramientas para el procesamiento del lenguaje oral y el escrito, tales como fonética y fonología, morfosintaxis, semántica léxica, pragmática y análisis del discurso. La factibilidad de estos desarrollos está sustentada con los avances actuales de los grupos de investigación a nivel nacional en esta materia.

Interfaces hombre-máquina: La investigación y desarrollo de interfaces hombre-máquina (IHM) son de relevancia para el diseño y desarrollo de sistemas que faciliten las tareas cotidianas a los usuarios de herramientas computacionales como las previstas en el CNCITL. Un ejemplo son los métodos para visualizar y recorrer grandes cantidades de información textual desde diferentes perspectivas relacionadas

al contexto del propio contenido de los documentos. Otro caso es el de las interfaces multimodales en las cuales los usuarios interactúan con la información haciendo uso simultáneo de dispositivos mecánicos (como el ratón) y de instrucciones orales que la computadora debe interpretar y combinar de manera adecuada para completar la tarea del usuario.

Los productos derivados de ésta línea de I+D se integran a las aplicaciones descritas en la sección 7 para facilitar el uso y el acceso a las fuentes de información con las que deben interactuar los usuarios.

Recuperación de información: La investigación y el desarrollo de sistemas para recuperación de información están directamente vinculados al diseño y producción de las herramientas para búsqueda inteligente de información descritas en la sección 7. Los procesos de estas herramientas suceden en diferentes momentos y bajo diferentes circunstancias, siendo los más visibles para los usuarios aquellos procesos que suceden en línea y en tiempo real. Por otro lado, algunos de los procesos más importantes como el modelado de la información almacenada en el sistema suceden fuera de línea; es en éstos procesos donde generalmente se aplican los métodos y herramientas más exhaustivas, como los de procesamiento del lenguaje natural.

Los productos del CNCITL en materia de recuperación de información comprenden desde una plataforma para búsqueda de información de propósito general, hasta herramientas de búsqueda especializadas sustentadas por lexicones y ontologías de dominio.

Las actividades de I+D relativas a estas tecnologías corresponden a las siguientes cuatro etapas de estos sistemas: Repositorios de información (almacenamiento, modelado, representación, e indexado de la información), motores de búsqueda (métodos de búsqueda, métodos para la expansión de la petición, métricas de similitud), generación de resultados (criterios de relevancia, clasificación, agrupamiento) y métodos de evaluación del sistema.

Extracción de información: La investigación y desarrollo en materia de extracción de información se relaciona directamente con las aplicaciones para el poblado automático de bases de datos a partir de fuentes de información no estructuradas (ver sección 7). Tal es el caso de la creación de directorios, lexicones y bases de conocimiento. Asimismo se relaciona con aplicaciones para la obtención de datos concretos a partir de grandes colecciones de información.

Esta área incluye procesos de búsqueda y recuperación de información en sus primeras etapas, generalmente a manera de recuperar documentos relevantes a ciertos contenidos. Una vez que se cuenta con los documentos que potencialmente contienen la información de interés se comienza con lo que propiamente es el proceso de extracción de información.

Las actividades de I+D relativas a estas tecnologías corresponden a las siguientes etapas de estos sistemas: clasificación de documentos relevantes, generación de patrones de extracción, desarrollo algoritmos de aprendizaje para identificación de *nuggets*, desarrollo de algoritmos para extracción de información, llenado de plantillas de información, poblado de bases de datos a partir de la información extraída.

En conjunto, las tecnologías mencionadas serán la base para el desarrollo de los productos y servicios propuestos en la sección 7 y orientados a las vertientes de búsqueda inteligente de información, diseminación selectiva de información y poblado automático de bases de datos. Estas herramientas permitirán crear tecnologías capaces de procesar y manejar grandes volúmenes de información, además de facilitar los mecanismos para la captación y el uso de la información oral y escrita de forma inteligente, cumpliendo así con las metas propuestas.

Actividades complementarias: Además de las actividades mencionadas, el personal de I+D participará activamente con la red multidisciplinaria de investigación básica y aplicada que promoverá el CNCITL. Dicha red será el resultado de las relaciones institucionales de cooperación y colaboración existentes actualmente entre los diferentes centros de investigación participantes en el CNCITL, así como con aquellos centros en el extranjero donde se han formado una gran parte de los investigadores mexicanos.

Para mantener informados a los miembros de la red se implantará un portal web desde el cual se pueda dar difusión a las diferentes actividades del Centro, así como a sus desarrollos y temas de interés.

Docencia y capacitación

Con objeto de aumentar la cantidad y calidad en la formación de profesionales de alta capacidad en las áreas afines al CNCITL, éste contribuirá a través de redes de trabajo con instituciones de educación superior nacionales e internacionales, a la propuesta, revisión y formulación de planes de estudio de alto nivel competitivo para su implantación dentro del territorio nacional. Asimismo promoverá estancias de capacitación y graduación que faciliten a las instituciones educativas la especialización de alumnos de pre y postgrado. Además del fortalecimiento de los profesionales existentes y de aquellos interesados en incursionar en las áreas mencionadas mediante programas educativos para la actualización continua.

Para alcanzar las metas en formación de recursos humanos, fomento a las TLIC en niveles desde pregrado hasta postgrado y la especialización de empresas, el CNCITL realizará las siguientes actividades:

- 1) Elaboración y recomendación de planes de estudio a las instituciones educativas.
- 2) Difusión de los beneficios y ventajas del desarrollo y uso de las TLIC en los sectores académico, productivo y de gobierno.
- 3) Capacitación continua y planes de especialización en TLIC tanto para las empresas usuarias como para las desarrolladoras de software.
- 4) Coordinación de una red multidisciplinaria de trabajo entre las instituciones que cooperan y colaboran actualmente en el CNCITL.

Vinculación empresarial

Las actividades están encaminadas a la vinculación con los sectores público, privado y social para el fomento de las tecnologías, recursos y productos desarrollados localmente; a la

comercialización de los servicios y productos del CNCITL; así como a la creación de alianzas estratégicas.

En este tenor se distinguen dos vertientes; por un lado se trata de la comercialización y difusión de las tecnologías desarrolladas en el CNCITL. Por el otro se buscarán alianzas estratégicas que permitan incluir las tecnologías desarrolladas en el CNCITL como parte de soluciones de fabricantes de software existentes en el mercado.

Comercialización: La comercialización de los productos y servicios desarrollados en el CNCITL requiere de la realización de un estudio prospectivo para la determinación de la demanda de productos y servicios tanto para el consumo nacional, como para el consumo en el extranjero. Los resultados arrojados por dicho estudio permitirán desarrollar de forma adecuada a las tendencias del mercado y a las tecnológicas, los productos y servicios creados en el Centro, focalizándose así en un mercado meta para cada tipo de herramienta o aplicación desarrollada.

Este departamento deberá colaborar de forma estrecha con el de alianzas estratégicas, ya que además de la comercialización directa a los usuarios de las TLIC, se prevé la generación de alianzas con empresas de diferentes industrias.

Alianzas estratégicas: El proceso de integración del CNCITL con la industria se considera desde tres perspectivas. Por un lado, viendo a la industria como un consumidor potencial, con requerimientos específicos de soluciones y el cuál puede verse involucrado en el proceso de nuevos productos y/o servicios. Por otro lado, la integración con la industria, en particular con la de TI, se plantea visualizando a las empresas desarrolladoras o productoras de software como clientes que requieren que el CNCITL les asesore o realice un proceso de transferencia tecnológica para que dichas empresas integren como parte de sus productos existentes nuevas funcionalidades o generen nuevos productos con base en las TLIC. Para ello el CNCITL debe desarrollar los mecanismos apropiados de transferencia tecnológica. Además se buscará establecer alianzas con empresas generadoras de información, tales como agencias periodísticas, revistas, editoriales, etc., de forma tal que los productos y servicios del CNCITL soporten las actividades de acceso y gestión de dicha información.

Finalmente, la integración con fábricas de software a las cuáles se les proveerá con los prototipos desarrollados por el CNCITL para que estas empresas desarrollen e implementen por cuenta propia productos y/o servicios derivados del conocimiento de las TLIC del CNCITL.

7. APLICACIONES A CORTO PLAZO

Los beneficios de contar con tecnologías del lenguaje y el aprovechamiento de sus recursos derivados son varios para la gestión de información, entre los que podemos encontrar:

Búsqueda inteligente de información

Permitirá a los usuarios de bases de datos documentales realizar búsquedas textuales en dichos repositorios de información utilizando frases y contexto, reduciendo así deficiencias de los buscadores actuales, como el gran número de resultados carentes de sentido a la necesidad de información y la necesidad de realizar manualmente múltiples búsquedas para satisfacer

dicha necesidad. Por ejemplo, para realizar la búsqueda rápida y precisa de tópicos en archivos de jurisprudencia, actuaciones en los archivos del ministerio público, o en bibliotecas digitales.

Distribución selectiva de información

Permitirá a los usuarios contar con agentes encargados de monitorear fuentes de información en línea. Dichos agentes mantienen el perfil de su usuario con las preferencias particulares de información, de manera que cuando detectan nueva información, envían una notificación al usuario con un resumen de dicha información y con los accesos necesarios para que pueda consultarla. De esta forma se puede contar con la información segundos posteriores a su publicación, no importa la diferencia horaria o geográfica.

Poblado automático de bases de datos

Permitirá a los usuarios obtener, a partir de textos, conjuntos de datos de forma automática, que luego son almacenados en bases de datos para la posterior generación de estadísticas, análisis de datos o gráficas. Un ejemplo es el prototipo del sistema TOPO [4] que tiene la capacidad de catalogar y analizar en pocos minutos aproximadamente 33,000 notas periodísticas sobre desastres naturales en México, extrayendo datos como el tipo de desastre, fechas, daños materiales, pérdidas humanas, etc., para almacenarlos en una base de datos. Estos sistemas pueden ser habilitados para trabajar en diferentes dominios de conocimiento.

8. DEMANDA ESPERADA

En virtud de que una gran cantidad de aplicaciones funcionan a partir del conocimiento, es previsible una alta perspectiva respecto a la demanda esperada por los productos y servicios del CNCITL. Por ejemplo, cabe mencionar proyectos de digitalización de la totalidad de documentación oficial (registro civil, catastro, registro vehicular, asuntos legales, archivo del ministerio público, expedientes en el sector salud, etc.) en diferentes estados y municipios de la república mexicana, dejando abiertos nichos de gran oportunidad para la aplicación directa de las tecnologías que se desarrollarán en el CNCITL. Otros proyectos incluyen, por ejemplo, la digitalización, conservación y acceso de colecciones documentales de alto valor histórico en conjunto con instituciones de investigación en ciencias sociales y el Archivo General de la Nación.

El sector financiero es otro claro ejemplo de las necesidades actuales de herramientas de alta tecnología. Este sector requiere de herramientas computacionales dotadas de inteligencia artificial para detectar patrones de actividad sospechosos, robo de identidades, y robo de información privada. Dichas herramientas son susceptibles de desarrollarse con base en las tecnologías del lenguaje y ofrecerse como servicios de inteligencia para la prevención y seguimiento de fraudes.

En el sector privado los ejemplos también abundan, como aplicaciones directas mediante alianzas con industrias afines a las actividades del CNCITL, por ejemplo, empresas que ofertan servicios de administración de almacenes documentales físicos y digitales. Otro ejemplo son las agencias de noticias, diarios y, en general, medios informativos, los cuales podrán ofrecer a sus suscriptores y público en general servicios personalizados de acceso a la información a través de los servicios de distribución selectiva de información del CNCITL. Por otro lado, el CNCITL puede desarrollar tecnologías que permita

posicionarse mejor en el mercado a las empresas que ofrecen servicios de creación de bases de datos.

9. PERTINENCIA DEL CNCITL

Un aspecto muy importante del CNCITL es el énfasis que le da a la vinculación con los sectores público, privado y social. El beneficio que puede obtenerse en el ámbito académico, empresarial, social y gubernamental es claro: contar en México con un campo propio de investigación y desarrollo científicamente consolidado y de calidad, en cuestiones de tecnologías de la información, permite resolver un sinnúmero de problemas ocasionados por la falta de métodos y control de datos que sean transformables en conocimiento. Por ejemplo, si se toma esto en cuenta, con la creación de una política inteligente de acceso a la información, se logra un avance importante para una interacción eficiente entre los sectores industriales y estatales, con necesidades concretas de conocimiento, con los sectores académicos, productores y gestores del mismo, y con la sociedad en general. Al establecer un sistema de procesamiento eficiente de intercambio de conocimiento entre las partes mencionadas, se pueden realizar acciones concretas con resultados óptimos que ayuden realmente a los diferentes estratos de la población, por ejemplo pequeñas y medianas empresas (PyMEs) con objetivos de comercio internacional que coadyuven al desarrollo económico y sustentable del país.

Otro aspecto de la propuesta que no debe pasar desapercibido es el de la formación de cuadros profesionales aptos de desenvolverse en el área, tanto en investigación como en desarrollo. El fomento a la investigación básica y aplicada, así como al desarrollo experimental, impacta directamente en la creación y puesta en marcha de un programa de formación de científicos y tecnólogos que labore eficaz y productivamente en las interdisciplinas de las tecnologías del lenguaje, la información y el conocimiento [5].

En concreto, al contar con una planta estructurada de investigación y desarrollo en todas sus vertientes junto con una base académica con formación sólida, el paso inmediato es explotar de un modo inteligente todos los aspectos relacionados con la captación y uso de información, lo que sin duda tendrá un impacto positivo en el crecimiento de otras áreas de conocimiento, como bioinformática, nanotecnología, vigilancia tecnológica, sistemas de redes y telecomunicaciones, ingeniería y gestión de conocimiento, creación de bibliotecas digitales, interacción hombre-máquina, procuración de justicia, lingüística forense, entre otras.

Cabe señalar que la creación de un centro de esta naturaleza fue propuesta recientemente al Consejo Nacional de Ciencia y Tecnología (CONACYT) por iniciativa del Instituto de Ingeniería de la UNAM, con la participación de grupos de investigación de diversas instituciones en el país, como son el Instituto Politécnico Nacional, el Instituto Nacional de Astrofísica, Óptica y Electrónica, la Benemérita Universidad Autónoma de Puebla, El Colegio de México, y de la misma UNAM. Actualmente esta propuesta, después de haber sido aprobada en su segunda fase, está en proceso de maduración y consolidación mediante las Redes Temáticas de Tecnologías de la Información propuestas por CONACYT.

10. AGRADECIMIENTOS

La presente propuesta surge en el contexto de la Convocatoria para Presentación de Ideas para la Realización de Megaproyectos de Investigación Científica o Tecnológica 2006, convocada por CONACYT (Proy. 54523), así como en el marco del Macroproyecto de Tecnologías para la Universidad de la Información y la Computación del Programa Transdisciplinario en Investigación y Desarrollo, UNAM. Fue realizada por iniciativa del Instituto de Ingeniería de la UNAM, con la participación de grupos de investigación de diversas instituciones en el país, como son el Instituto Politécnico Nacional, el Instituto Nacional de Astrofísica, Óptica y Electrónica, la Benemérita Universidad Autónoma de Puebla, El Colegio de México, y de la misma UNAM.

11. REFERENCIAS

- [1] L. Orozco, "Sistemas y tecnologías de información en las micro y pequeñas empresas", en *Razón y Palabra*, núm 39, junio-julio 2004.
- [2] A. Gelbukh y G. Sidorov, *Procesamiento automático del español con enfoques en recursos léxicos grandes*, México: Instituto Politécnico Nacional, 2006.
- [3] CONACYT, *Programa Especial de Ciencia y Tecnología 2001-2006*, México: Consejo Nacional de Ciencia y Tecnología, 2001.
- [4] A. Téllez, M. Montes, O. Fuentes y L. Villaseñor, "Clasificación automática de textos de desastres naturales en México", en *Memorias del 10º Congreso Internacional de Investigación en Ciencias Computacionales (CIICC 2003)*, 2003.
- [5] G. Sierra y A. Medina, "Ingeniería Lingüística: Interdisciplina aplicada al desarrollo humano" en *Memorias del Primer Congreso de la Academia de Ingeniería*, Academia de Ingeniería, México, 2003.