Primera Asociación para el procesamiento del lenguaje natural en América Latina: problemas y perspectivas*

Alexander Gelbukh

Centro de Investigación en Computación,
Instituto Politécnico Nacional, CP 07738, México.
gelbukh@cic.ipn.mx, www.cic.ipn.mx/~ gelbukh

Resumen La formación de una asociación para el procesamiento del lenguaje natural en un país latinoamericano tiene una problemática específica. Una de sus tareas importantes es la interacción con las organizaciones gubernamentales para ayudarles a adecuar su actitud hacia esta ciencia. Una decisión crítica es la del financiamiento de la asociación; discutimos las consecuencias del rechazo del financiamiento alguno. La discusión se basa en el proceso de la formación de tal asociación en México.

1 Introducción

En las últimas décadas, el procesamiento de lenguaje natural (PLN) se desarrolló en una ciencia madura y bien organizada. Existen en el mundo grandes institutos y laboratorios dedicados a esta ciencia, específicamente en los EE.UU. y en Europa, donde su desarrollo fue impulsado fuertemente por la unión de Europa.

Un indicador de la madurez de la ciencia es el florecimiento de las asociaciones profesionales. La asociación mundial –ACL [1]– tiene una desarrollada estructura, con los capítulos en Europa, Norteamérica y varios grupos de intereses (*SIG* en inglés). En muchos países, sobre todo en Europa, existen asociaciones o sociedades nacionales. Por ejemplo, sólo la SEPLN [2] en España une a más de 800 miembros.

La situación en América Latina es diferente. Hay pocos investigadores que trabajan en el PLN. En consecuencia, prácticamente no existe la infraestructura correspondiente. Según nuestro conocimiento, la única organización completamente dedicada al estudio computacional de lenguaje en América Latina es el Laboratorio de lenguaje natural y procesamiento de texto del IPN, México [3].

Sin embargo, el número de especialistas y estudiantes interesados en el PLN en los países latinoamericanos, sobre todo en México, Cuba, Uruguay, Chile, Argentina llegó a una cifra suficiente para juntar sus esfuerzos en las cuestiones del interés común.

Una de las formas organizacionales de tal colaboración son las asociaciones profesionales.

En este artículo, damos una discusión de la necesidad de organización de la asociación para el PLN en México (con la esperanza que los mismos argumentos se aplican a otros países latinoamericanos) y discutimos los problemas que surgieron en la organización de la primera asociación de este tipo en América Latina, a saber, la Asociación Mexicana para el Procesamiento de Lenguaje Natural (AMPLN) [4].

2 Razones para la organización nacional

Existe un interés mutuo para la participación de los especialistas latinoamericanos en las asociaciones profesionales, tales como la ACL (asociación mundial) y SEPLN (España). En cuanto a la ACL, el marco más natural sería la afiliación con la NAACL, su capítulo norteamericano. En cuanto a la SEPLN, se propuso su extensión para formar la Sociedad Iberoamericana (SIAPLN).

Hay cuatro modelos de tal participación:

- *Participación individual*: cada especialista puede hacerse miembro de una asociación ya existente, por ejemplo, de la SEPLN.
- *Capítulos* regionales o nacionales de una asociación existente.
- *Asociación regional* como SIAPLN (tomando en cuenta el número reducido de los miembros potenciales en cada país).
- *Asociación nacional*; las asociaciones nacionales se pueden organizar después en una fe-

* Expresamos los más cordiales agradecimientos a Eduard Hovy, Ted Pedersen, Dina Wonsever y Aurelio López-López por sus útiles consejos y discusión.

El trabajo fue parcialmente financiado por el CONACyT, CGEPI-IPN y SNI, México.

Proc. SLPLT-2, the 2nd International Workshop on Spanish Language Processing and Language Technologies, Jaén, Spain, September 14–15, 2001, ISBN 84-8454-095-2, Editorial Club Universitario, pp. 35–40.

deración regional y también afiliarse a las organizaciones grandes existentes.

Creemos que el último modelo es más adecuado, por las siguientes razones.

- Puede promover la colaboración local.
- No choca con la mentalidad latinoamericana[1].
- Una función importante de la asociación es interacción con los organismos gubernamentales que toman las decisiones financieras y administrativas sobre el desarrollo de la ciencia. Creemos que con más probabilidad harán caso a una organización nacional que a una internacional o regional.

La última consideración determinó nuestra decisión para formar la asociación nacional.

Además, para el diálogo productivo con las instancias oficiales, se necesita organización del mismo nivel que la instancia correspondiente. Esto implica la necesidad de integrar las asociaciones nacionales en una federación regional (de América Latina o de Iberoamérica).

Es decir, nuestra decisión implica una conjunción de todas las cuatro opciones.

3 La ciencia y la sociedad

Uno de los problemas más importantes para el desarrollo del PLN en América Latina es la insuficiencia de actividad, de especialistas, profesores y estudiantes.

Se forma un círculo vicioso: hay pocos estudiantes, entonces, hay pocos profesores; con esto, la industria y el gobierno no soportan el desarrollo del PLN; entonces, los estudiantes no esperan, cuando egresarán, obtener un trabajo relacionado con la especialidad obtenida. Por eso hay pocos estudiantes.

¿Cómo podemos romper este círculo? Desgraciadamente, tenemos más preguntas que respuestas.

3.1 La industria

¿Podemos convencer a la industria de que somos útiles? ¿Qué productos podemos ofrecer? ¿Podrán algún día nuestros estudiantes obtener buenos trabajos en la industria?

¿Por qué desarrollar el software lingüístico en nuestros países es mejor que comprarlo de los países más desarrollados? Desarrollo lingüístico serio es caro, requiere de los recursos léxicos costosos y/o mucha labor a mano de buenos lexicógrafos que también es cara.

La idea de desarrollar las herramientas para los lenguajes indígenas (aymará, quechua, etc.) no es muy prometedora para interesar a la industria en nuestros países.

3.2 La academia y el gobierno

Por otro lado, la ciencia no necesariamente debe crear productos inmediatamente vendibles. Pocos astrónomos o filósofos trabajan en la industria. Sin embargo, el desarrollo de estas ciencias es financiado por los gobiernos de la mayoría de nuestros países. Hay suficientes plazas para los profesores, y los estudiantes están bastante seguros en obtener trabajo en alguna universidad.

Claramente, un país puede optar por no gastar presupuesto en astronomía, obteniendo la información necesaria (si un día la necesitase) de los libros publicados en el mundo.

Sin embargo, la integración del país en el proceso científico mundial es una aportación muy importante en el prestigio del país (igual al deporte), un impulso a su cultura y educación, así como una condición necesaria para su integración en la humanidad.

En cuanto a las ciencias tradicionales (tales como física, biología, química), nuestros gobiernos están concientes de la necesidad del desarrollo de la ciencia. Pero en cuanto al PLN (y en general la ciencia de la computación), todavía tenemos que convencerlos en que es una ciencia no menos importante que la astronomía.

En México el organismo que determina el desarrollo de la ciencia es el Consejo Nacional de Ciencia y Tecnología (CONACyT). Éste mantiene al Sistema Nacional de Investigadores (SNI). Tanto los ingresos personales como el financiamiento estatal de los proyectos de investigación dependen fuertemente de la membresía (y el nivel) de los investigadores en el SNI.

En las reglas del SNI hay ciertos factores que dificultan el desarrollo del PLN en el ambiente académico, sobre todo los siguientes:

- Falta de reconocimiento del PLN como una ciencia interdisciplinaria.
- Inercia de los criterios para medir los logros científicos.

Una tarea importante de la asociación será ayudar a los organismos gubernamentales (co-

[1] En nuestra experiencia, algunos colegas rechazan –explícita, implícita o inconscientemente– la idea de la expansión directa de la antigua metrópoli o el poderoso vecino norteño a lo que consideran un asunto nacional. Sin discutir la relevancia de semejantes consideraciones en los asuntos de la ciencia, creemos que estos sentimientos se deben respetar y tomar en cuenta.

mo CONACyT) e institucionales (que en muchos casos copian los criterios del SNI) a efectuar las adecuaciones necesarias.

En cuanto al primer punto, el problema es que el PLN no cabe en la clasificación tradicional usada por el SNI.

Considerado como una rama de las humanidades, en muchos casos causa el rechazo de parte de los lingüistas pues la mayoría de los métodos que usamos (digamos, métodos estadísticos) se ven muy simplificados o simplemente incorrectos si consideran como investigación puramente lingüística (por la cual en los países latinoamericanos se entiende principalmente el estudio de las lenguas indígenas).

Por otro lado, considerado como una rama de la ciencia de la computación, también causa rechazo porque una gran parte de nuestra investigación se ve como la de humanidades. También causa rechazo por los criterios formales que explicamos a continuación (ya hablando del segundo punto –inercia de los criterios).

El primer requisito para la aceptación de un investigador al SNI (área de Ingeniería) es publicar los artículos en las revistas de alto prestigio, las cuales se definen por el SNI como aquellas indexadas por el *Science Citation Index* (SCI) de la compañía estadounidense ISI [5]. Sin embargo, el ISI publica unas tres decenas de índices, de los cuales el SCI es sólo uno [6]. Por la clasificación del ISI, las revistas relacionadas con el PLN pertenecen al *Social Science Citation Index* (SSCI) y no al SCI. Entonces, bajo el cumplimiento estricto de las reglas vigentes, los especialistas en el PLN nunca jamás tendrían la posibilidad de entrar en el SNI.

Otro problema del uso indiscreto de los criterios del ISI por el SNI es que el ISI recientemente excluyó de sus índices la mayoría de las revistas que se publican en los idiomas distintos al inglés, motivándolo con las razones económicas (no tienen traductores) [7]. Lo que afecta a los investigadores que publican en español.

La idea de considerar sólo las publicaciones en las revistas (totalmente ignorando las publicaciones en las memorias de los congresos) en general se considera anacrónica por muchos de nuestros colegas. En el área de computación en general, y específicamente del PLN, hay pocas revistas, y muchos trabajos muy importantes se publican en las memorias de los congresos internacionales [8]. Se deberían tomar en cuenta como publicaciones científicas.

Sin embargo, este problema se ve sumamente complejo. Las revistas son estables, y se tiene un mecanismo bien desarrollado para medir la importancia de una revista (los índices del ISI). A diferencia, no existe un mecanismo simple y seguro para que un no especialista en el área (un evaluador del SNI) pueda evaluar la importancia de un congreso.

Otra vez, una tarea de la Asociación sería proponer (junto con los colegas de otras áreas de la computación) a las organizaciones correspondientes las posibles soluciones a este problema que consideramos crítico para el desarrollo del PLN en nuestro país.

4 Otras actividades

Por supuesto, la AMPLN desempeñará otras actividades que las que se discutan en la sección anterior, más tradicionales para una asociación de esta índole. En general, sus objetivos son:

- *Representar* a la comunidad de los especialistas en PLN; proporcionarle la «voz colectiva».
- *Promover* la interacción e intercambio de las ideas, herramientas y resultados.
- *Difundir* los logros y la importancia del PLN en la sociedad.

Específicamente, el segundo objetivo (la promoción de la colaboración) se puede dividir en la colaboración interna y externa.

4.1 Colaboración interna

Un objetivo importante de la Asociación es impulsar la colaboración entre los pocos especialistas que trabajan en el país, es decir, entre sus miembros. Las siguientes actividades soportan principalmente este objetivo.

- Mantener un sitio web [4] con la información acerca de las organizaciones, personas y recursos en México y en el extranjero relevantes para la comunidad mexicana del PLN.
- Organizar periódicamente un mini-congreso o seminario donde los miembros conocerán el trabajo actual uno de otro.
- Apoyar a la organización de congresos en México (o en el extranjero) que involucran el PLN, entre éstos, la serie CICLing [9].

4.2 Colaboración externa

La colaboración externa de la Asociación se puede dividir en varias dimensiones: 1) colaboración o afiliación con las asociaciones semejantes en extranjeras o internacionales, 2) colaboración con las asociaciones nacionales en las

áreas afines, 3) interacción con las instituciones y organizaciones nacionales.

Entre estas tareas hay los siguientes:

- Promover las publicaciones de los miembros a través del sitio web de la asociación.
- Apoyar a las labores de editoriales de las revistas que involucran los temas del PLN.
- Proporcionar a las organizaciones nacionales y extranjeras la información sobre el estado del PLN en el país.
- Difundir el conocimiento sobre la importancia y las realidades actuales de la ciencia del PLN, a las instituciones nacionales mediante las cartas, aclaraciones, recomendaciones a las personas y proyectos etc.
- Colaborar con las organizaciones y asociaciones nacionales (p.ej. la Academia de Ciencias), extranjeras (p.ej. SEPLN) o internacionales (p.ej. SIAPLN, ACL).

En cuanto al último punto, una posible forma de colaboración con ACL sería la distribución en América Latina de la revista *Computational Linguistics* con un significativo descuento. Se pueden considerar los descuentos para los miembros de AMPLN por parte de la ACL [8].

5 El dilema de financiamiento y sus consecuencias

El punto clave en el funcionamiento de cualquiera organización es su fuente de financiamiento. El dilema principal es: manejar o no manejar sus propias finanzas.

Tomando en cuenta el muy bajo número de los miembros potenciales, así como bajos ingresos promedios en nuestro país, creemos que no podremos lograr el ingreso anual suficiente para el mantenimiento de las actividades que lo requieren. Además, la burocracia necesaria para el manejo de los recursos no vale la pena cuando éstos son tan pocos.

Entonces, tomamos la decisión crítica: por el momento no mantener los recursos financieros.

Esto, a su vez, implica otras complicaciones en el funcionamiento de la Asociación.

Es obvio que limita las posibilidades de la Asociación en las funciones básicas tales como organización de congresos o mantenimiento de una revista.

Pero también crea un nuevo problema: el de la membresía. En una asociación con pago anual de inscripción, la pregunta «quién puede ser miembro» se resuelve automáticamente: sólo las personas interesadas en sus actividades y objetivos pagan aceptan el pago anual.

Pero sin este factor regulador y considerando que la membresía en la asociación representa ciertas ventajas y cierto honor, se necesita definir precisamente el objeto de la asociación, qué tiene derecho a pertenecer a ésta, y quién toma las decisiones sobre la admisión.

Específicamente, si no se definen bien las fronteras de la Asociación (¿pueden entrar lingüistas puros? ¿y maestros de lenguas extranjeras? ¿y los desarrolladores de bases de datos documentales?) y tomando en cuenta la organización democrática (los miembros votan por las decisiones importantes), los objetivos de la Asociación se pueden transformar y la Asociación se puede dejar de ser una asociación profesional útil para sus miembros específicos.

5.1 Definiciones básicas

Antes de nada, es difícil definir el objeto de la Asociación. Incluso no hay consenso sobre cual de los términos: *lingüística computacional*, *procesamiento* (*tratamiento*) *del lenguaje natural*, *lingüística aplicada*[2], etc. es más genérico y cuál corresponde mejor a la ciencia que representamos. Por ejemplo, la asociación mundial (ACL) usa el primer término, la española (SEPLN) el segundo, y la francesa (ATALA [11]) el tercero, aunque se trata de la misma ciencia.

Nuestra selección del término *procesamiento del lenguaje natural* se basa en parte en el sentimiento que este término es el más preciso para la ciencia que representamos y en parte en la selección de la SEPLN.

En la reglamentación de la Asociación se definen los siguientes términos:

- Por el *lenguaje natural* se entiende cualquier lenguaje humano (p.ej. español, inglés etc.) en sus aspectos lingüísticos.

  No se considera *lenguaje natural* el texto o voz considerados fuera del aspecto lingüístico (p.ej. considerados en el contexto de transmisión de las páginas web por la red, formatos de archivos de sonido etc., a menos que se involucran aspectos lingüísticos).

[2] Parece que la definición más común del termino «lingüística aplicada» implica las aplicaciones ajenas a la computacional. Por ejemplo, así entiende este término la ACLA/CAAL –la Asociación Canadiense para la Lingüística Aplicada [10].

- Por el PLN se entienden las ramas de la ciencia de computación, matemáticas e ingeniería que involucran el estudio o aplicaciones del procesamiento *automático* del *lenguaje natural* por *computadora*, incluidos los estudios lingüísticos cuyos resultados son directamente útiles para el desarrollo de los programas computacionales para el PLN (aunque sus autores no desarrollan tales programas).

  En esta interpretación, PLN se entiende como el término más amplio que describe *todo lo útil para la creación de los programas que procesan el lenguaje natural*. Comprende la lingüística computacional, lingüística aplicada (las aplicaciones que involucran realización computacional), ingeniería de lenguaje, tecnologías de lenguaje, entre otras áreas.

  No se consideran PLN:

  - las diversas ramas de la lingüística pura o cognitiva que estudian el lenguaje humano sin relación con las computadoras ni con métodos computacionales;
  - ni las diversas ramas o aplicaciones de la ciencia de computación o ingeniería que, aunque involucran el procesamiento automático de voz o texto (p.ej. transmisión, impresión etc.), no tienen que ver con el lenguaje natural;

  aunque se reconoce que cada de estas ramas de ciencia es una ciencia importante y una de las bases fundamentales del PLN;

  - ni las aplicaciones de los métodos desarrollados por el PLN (p.ej. gramáticas formales), fuera del contexto de PLN.

5.2 Estructura y membresía

Para aplicar flexiblemente las nociones de la sesión anterior a la aceptación de nuevos miembros, resultó necesario establecer una comisión de membresía de tres miembros, dos de los cuales serían (para no inflar la burocracia) el Presidente y Vicepresidente. Entonces, la estructura administrativa se fijó como sigue: presidente, vicepresidente, secretario, miembro vocal de la Comisión de Membresía, tesorero (este puesto no se establece hasta el momento cuando se establezcan algunas fuentes de financiamiento).

Como una secuencia de la consideración del ingreso de nuevos miembros por una comisión se hizo posible establecer varios niveles de membresía:

- *Miembro regular*. Son los especialistas en el PLN, lo que comprueban con sus publicaciones. Sólo éstos toman parte en votaciones.
- *Miembro asociado*. Son las personas interesadas en el PLN (estudiantes, empresarios, etc.) pero no especialistas.
- *Miembro honorario*. Ser miembro honorario de la AMPLN representa un gran honor que se otorga en reconocimiento de los meritos excepcionales, usualmente a los extranjeros.

Una ventaja del proceso no automático de la admisión es que la membresía en la AMPLN puede considerarse como un honor o distinción.

Para elegir a los funcionarios de la Asociación por primera vez, el fundador invitó a ser miembros a los especialistas nacionales que cumplen con los requisitos establecidos. Hasta el momento, se invitó a 19 miembros regulares y a 9 miembros asociados (estudiantes).

6 Conclusiones

La formación de una asociación para el PLN en un país latinoamericano tiene una problemática específica:

- La justificación de su formación;
- Los problemas de financiamiento;
- La definición de las fronteras de la asociación y de la admisión de los miembros (estos problemas surgen de la decisión de imponer cuotas de inscripción);
- Los problemas de las actividades (que no requieran de financiamiento).

También el objetivo importante de tal asociación en un país latinoamericano puede incluir la interacción con los organismos gubernamentales apoyando la adecuación de sus criterios y acciones a las realidades de nuestra ciencia.

En este artículo hemos presentado nuestras reflexiones y la experiencia de la organización de tal asociación en México que está en su fase inicial. Esperamos que nuestra experiencia será útil para los compañeros de otros países para la organización de las asociaciones nacionales, las cuales esperamos unir en una federación regional –SIAPLN.

Referencias

1. www.aclweb.org
2. www.sepln.org
3. www.cic.ipn.mx/Investigacion/ltexto.html
4. Dirección provisional: www.CICLing.org/ampln
5. www.conacyt.mx/sni/sni0027.html

6. sunweb.isinet.com/isi/journals/index.html
7. sunweb.isinet.com/isi/hot/essays/selectionofmaterialforcoverage/199701.html
8. Eduard Hovy, Presidente de la ACL; comunicación privada.
9. www.CICLing.org
10. www.aclacaal.org
11. www.atala.org