Aplicación de técnicas de PLN en sistema de donaciones

Martín Santillán Cooper[1], Francisco Serrano[2], Marcelo Armentano[1], Silvia Schiaffino[1], María Rosa Dos Reis[2]y Moisés Bueno[2]

[1]ISISTAN (CONICET-UNCPBA), Campus Universitario, Tandil, Argentina.
[2] Facultad de Ciencias Exactas, UNCPBA, Campus Universitario, Tandil

Resumen En este artículo se decribe el enfoque basado en Procesamiento de Lenguaje Natural adoptado en el sitio AYUD@RG para poder detectar de manera satisfactoria las donaciones que las personas quieren realizar a partir de texto libre. El objetivo de AYUD@RG es poner en contacto a los donantes y a aquellas ONGs que necesitan los recursos en cuestión. Los principales desafíos los constituyen las caracteríasticas propias el lenguaje natural, como la ambigüedad en los términos, los sinónimos, los errores ortográficos. Resultados preliminares demuestran la viabilidad del enfoque adopatado.

1. Introducción

AYUD@ARG[1] es una solución integral al trabajo en redes sociales de las instituciones que vinculan la oferta y demanda de recursos. Este sistema es el puente entre la necesidad de ciertos recursos que tienen algunas personas y la ayuda que brindan las instituciones con vocación social. La página Web de AYUD@RG permite que ONGs u otro tipo de usuarios puedan gestionar los recursos que requieren para sus proyectos. Por un lado, el sistema permite realizar la donación de recursos, y por otro, permite que dichas ONGs pidan recursos. AYUD@RG es un servicio de la Asociación Civil Proyecto KOINONIA[2].

Uno de los aspectos claves de AYUD@RG consiste en la detección y compatibilización de recursos a donar u ofrecer a través del sitio Web a partir de texto libre. En este trabajo se detalla el proceso realizado para lograr este objetivo utilizando técnicas de Procesamiento de Lenguaje Natural. Las tareas que se llevaron a cabo fueron las siguientes: reconocimiento de sinónimos, reconocimiento de hipónimos e hiperónimos, y desambiguación.

El resto del artículo se organiza de la siguiente manera. En la sección 2 se describe en detalle el problema resuelto. Luego, en la sección 3 se detalla el trabajo realizado. En la sección 4 se menciona un ejemplo de un caso de estudio realizado. Finalmente, en la sección 5 presentamos las conclusiones.

2. Definición del problema a resolver

El problema central consistió en permitir que la página web del Proyecto Koinonia reconozca de forma óptima los recursos que los usuarios piden o donan. Para ello es

[1] http://www.proyectokoinonia.org.ar/ayudarg/
[2] http://www.proyectokoinonia.org.ar

necesario resolver los problemas de origen semántico que debe solucionar cualquier sistema que busca hacer uso, en alguna de sus partes, del lenguaje natural. Los problemas a resolver fueron:

- Reconocimiento de sinónimos: los recursos a donar y aquellos que son pedidos pueden especificarse con diferentes palabras que tengan el mismo significado semántico. Es necesario implementar un mecanismo que permita que los sinónimos sean detectados e identificados de tal forma que aquellos recursos representados por los usuarios con dos palabras diferentes, sean hechos persistentes en la base de datos como un único recurso.
- Reconocimiento de hipónimos e hiperónimos: los hipónimos e hiperónimos son, al igual que los sinónimos, relaciones semánticas que se dan entre las palabras. Este tipo de relaciones semánticas representan para el Proyecto Koinonia, una interesante posibilidad a tener en cuenta para mejorar el sistema de donaciones. La hiponimia es la relación semántica en que el significado de una palabra (ej. verano) se encuentra incluido en otro de mayor amplitud (ej. estación), su hiperónimo.
- Desambiguación: al incorporar la sinonimia al sistema de donaciones se genera un problema adicional a resolver, que es la desambiguación. Esto se debe a que, al mismo tiempo que varias palabras pueden pertenecer al mismo conjunto de sinónimos, una palabra puede pertenecer a varios conjuntos de sinónimos (polisemia). Por lo tanto, es necesario implementar una etapa de desambiguación que permita al usuario decidir cuál es efectivamente el recurso que busca donar o pedir.

3. Enfoque Propuesto

Para desarrollar el trabajo se estudiaron en primer lugar diferentes proyectos existentes relacionados a los problemas planteados en la sección anterior. Se decidió utilizar WordNet [2], que es una base de datos organizada bajo la noción de synset, entre los que se mantienen relaciones semánticas. Un synset (del inglés synonym sets) es un conjunto de sinónimos que pueden estar relacionados entre ellos a través de hiponimia, meronimia, etc. Luego de un análisis de implementaciones de WordNet para el idioma castellano, se decidió utilizar Multilingual Central Repository (MCR) [1], por ser de libre acceso y por contar con varios años de desarrollo.

En la Figura 1 se detalla el proceso seguido desde que un usuario indica lo que quiere donar hasta que el sistema identifica el recurso. El proceso de búsqueda de coincidencias entre donaciones y pedidos está dividido en varias etapas, en donde inicialmente se cuenta con el texto libre que es ingresado por los usuarios en el sitio web del Proyecto Koinonía. Una de las etapas consiste en someter dicho texto libre a un detector de recursos [3], que consiste en una herramienta que tiene como entrada texto libre. El resultado es un listado de los bienes detectados. La única restricción que dicha herramienta presenta, es que el texto libre debe estar exento de errores ortográficos, por lo que como previamente a la detección de los recursos, el texto libre debe atravesar una etapa de validación ortográfica. Dicha etapa se lleva a cabo mediante la utilización de una herramienta libre llamada LanguageTool[3], en su versión para el lenguaje Java.

Una vez obtenidos los recursos a partir del texto libre, se los utiliza en el acceso a la base de datos del MCR para obtener los synsets asociados a dichos recursos. A partir de ese instante, cada palabra deja de estar identificada por el texto plano, sino por un

[3] https://languagetool.org/

identificador de synset. Aquí surge el problema de polisemia, dado que un recurso puede pertenecer a varios synsets. Los identificadores se utilizarán posteriormente para buscar las coincidencias, ya que tanto “manta” como “colcha” tienen el mismo identificador, al pertenecer al mismo conjunto de sinónimos. Una de las cuestiones que surgieron en torno a esta etapa fue la presencia de ambigüedades al momento de mapear cada palabra con su correspondiente synset. Por ejemplo “manta” puede referirse tanto a “colcha” como a “mantarraya”. Si bien el contexto del servicio indica que por sentido común el synset asociado debería ser aquél en donde también está presente la palabra “colcha”, es necesario que el usuario desambigüe los recursos ingresados al seleccionar el synset al cual el recurso se refiere. De esta manera queda cerrado el flujo de la información que inicia con la escritura del texto libre por parte del usuario, resultando finalmente en el registro de los recursos donados/ofrecidos que son identificados a través del synset al cual pertenece cada recurso.

Una vez desambiguados los recursos detectados, es decir, al haber mapeado cada recurso con su correspondiente identificador de synset (con ayuda de la intervención del usuario), se los almacena en la base de datos existente del sitio web de Koinonía. De esta manera ya es posible efectuar la búsqueda de coincidencias en caso de que se quiera donar/pedir un determinado bien. Adicionalmente, al registrarse la desambiguación del usuario para cada recurso detectado, es posible almacenar un ranking de synsets escogidos para cada recurso. Este ranking se irá perfeccionando conforme vaya incrementándose el uso del sitio web, para que así posteriormente se puedan efectuar recomendaciones al momento de desambiguar. A modo de ejemplo, si muchos usuarios siempre desambiguaron el recurso “manta” con el mismo synset (el correspondiente a “colcha”), dicho identificador de synset estará calificado como el más elegido para la palabra “manta”.

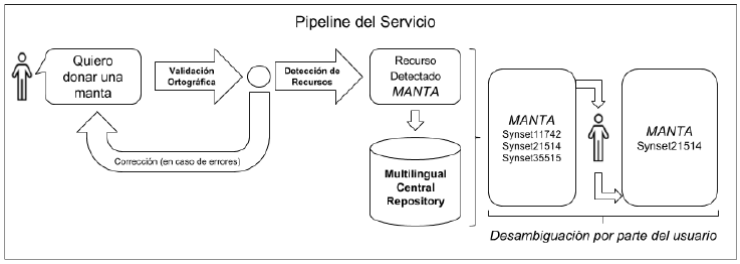

Figura 1. Procesamiento realizado

4. Caso de estudio

En esta sección, describimos tres casos de estudios mediante diferentes frases utilizadas por las personas interesadas en donar uno o más recursos. La Tabla 1 presenta los resultados obtenidos del proceso de PLN desarrollado. En el caso de aquellos recursos para los cuales es necesaria una desambiguación por parte del usuario, se muestran los diferentes synsets en los que se encuentra el recurso detectado de forma tal que el usuario puede seleccionar cuál de todas las acepciones posibles de la palabra corresponde al

recurso disponible para ser donado. Notar que para la primera consulta, la palabra “plaza" no fue detectada como recurso, como era de esperar. Para el tercer caso de estudio presentado, al obtenerse un único synset para la palabra “calesita", la misma es identificada automáticamente como recurso y no se requiere de la intervención del usuario para desambiguar.

Frase utilizada	Recursos detectados
"Quiero donar una manta y una cama de una plaza"	MANTA:58337:alfombra, alfombrilla, moqueta, tapete, MANTA:90954:holgazán, badana, perezoso, flojo MANTA:48854:colcha, frazada MANTA:26940:manta_ray, mantarraya CAMA:48456:lecho CAMA:48501:armazón_de_cama, cabezal, marco_de_cama CAMA:48473:lecho
"Tengo para ofrecer leche y manteca"	MANTECA:78414:mantequilla MANTECA:77265:lardo, manteca_de_cerdo MANTECA:77267:manteca_de_hojaldre CREMA:52291:emoliente, ungüento CREMA:72354:diéresis, trema, umlaut, metafonía, dialefa CREMA:78406:nata, crema_de_leche CREMA:76875:natilla, crema_pastelera, natillas CREMA:56724:salve, ungüento, bálsamo, pomada CREMA:63130:beige, beis CREMA:81181:la_crema
"Tengo para donar una calesita a reparar"	CALESITA:50300:caballitos, carrusel, tiovivo

Cuadro 1. Casos de estudio

5. Conclusiones

En este trabajo se describió el enfoque adoptado por el sitio AYUD@RG para detectar de manera satisfactoria los recursos donados por diferentes personas o instituciones. Consideramos, que si bien en enfoque adoptado no es novedoso en cuanto al uso de herramientas de Procesamiento de Lenguaje Natural, constituye un ejemplo de aplicación que vale la pena destacar. El servicio desarrollado ha sido integrado al sitio del proyecto Koinonia, sin embargo no se han realizado aún evaluaciones exhaustivas, y se encuentra aún en estado de prototipo.

Referencias

1. Gonzalez-Agirre, A., Rigau, G.: Construcción de una base de conocimiento léxico multilingüe de amplia cobertura: Multilingual central repository. Linguamática 5(1), 13–28 (2013)
2. Miller, G.A., Beckwith, R., Fellbaum, C., Gross, D., Miller, K.: Introduction to wordnet: An on-line lexical database. International Journal of Lexicography 3(4), 235–244 (1990)
3. Varona, B.: Un enfoque híbrido para la detección automática de recursos en textos cortos. Trabajo final de grado, Facultad de Ciencias Exactas, UNCPBA (2018)