Revista Digital Universitaria
1 de mayo 2012 • Volumen 13 Número 5 • ISSN: 1067-6079

ARTICULO

PROCESAMIENTO DE LENGUAJE NATURAL EN LA UNIVERSIDAD DE LA REPÚBLICA

Dina Wonsever, Marisa Malcuori y Aiala Rosá

1 de mayo 2012 • Volumen 13 Número 5 • ISSN: 1067-60710

Procesamiento de Lenguaje Natural en la Universidad de la República

Resumen

En este artículo se presentan algunos resultados del trabajo de investigación interdisciplinario llevado adelante por informáticos y lingüistas en la Universidad de la República. Se describen las líneas generales de la investigación, orientada hacia el procesamiento automático de textos, en particular, la propuesta de un modelo de discurso organizado en varios módulos capaces de interactuar entre sí.

Palabras clave: Procesamiento de Lenguaje Natural, Discurso, Eventos, Enunciación, Sintaxis

Abstract

This paper presents some results of interdisciplinary research carried out by computer scientists and linguists at the Universidad de la República. We describe the outline of the research in automatic text processing, in particular, the proposal of a discourse model organized in several modules that can interact with each other.

Keywords: Natural Language Processing, Discourse, Events, Enunciation, Syntax

Introducción

El área de procesamiento de lenguaje natural es por su naturaleza un espacio interdisciplinario entre la Informática y la Lingüística. La vinculación corresponde a una confluencia en el objeto de estudio, el lenguaje humano, ya sea con el propósito de conocer a fondo su estructura y funcionamiento, como con el de construir aplicaciones informáticas con capacidad de realizar comprensión y extraer información de enunciados lingüísticos.

En la Universidad de la República este espacio interdisciplinario se ha concretado en el trabajo conjunto de dos grupos académicos: el Grupo de Procesamiento de Lenguaje Natural (GPLN, www.fing.edu.uy/inco/grupos/pln) del Instituto de Computación de la Facultad de Ingeniería (www.fing.edu.uy/inco) y el Departamento de Teoría del Lenguaje y Lingüística General (DTTLG, http://www.fhuce.edu.uy/index.php?option=com_content&view=article&id=145%3Adepto-de-teoria-del-lenguaje-y-lingueistica-general&catid=70%3Ainstituto-de-lingueistica&Itemid=92), de la Facultad de Humanidades y Ciencias de la Educación (http://www.fhuce.edu.uy). Más de diez años de trabajo conjunto se han concretado en varios proyectos de análisis de textos: modelo de discurso, reconocimiento de proposiciones, reconocimiento de eventos y expresiones temporales y otros.

El hecho de procesar textos en forma automática, cualquiera sea la finalidad con que se realiza tal tarea (extraer información, responder preguntas, realizar resúmenes y traducir), implica enfrentarse con ciertas propiedades del discurso que es necesario aprehender.

Nuestra línea de trabajo ha sido la de adoptar una estructura modular para dar cuenta de esas propiedades complejas, expresándolas mediante el análisis en distintos ejes o módulos independientes, capaces, sin embargo, de interactuar entre sí. Esta estructura, si bien no aporta en principio una visión holística del discurso, permite, sin embargo, trabajar independientemente en cada eje, al mismo tiempo que habilita la incorporación de otros nuevos, a medida que se vayan desarrollando.

Los ejes propuestos son los siguientes: Enunciación/Eventos-Factividad/ Temporalidad/Estructura retórica. A estos cuatro ejes se agregan dos más de carácter estructural: Sintaxis/Estructura textual (párrafo, sección, título, etcétera). En las secciones siguientes presentamos un panorama del trabajo realizado en algunos de estos módulos.

Los módulos Eventos-Factividad, Enunciación y Sintaxis

El módulo Eventos-Factividad

En el eje Eventos-Factividad, los elementos pertinentes que es necesario detectar y anotar, son los eventos a los que se hace referencia en los textos.

¿Qué entendemos por evento? Definimos un evento como cualquier tipo de situación o acontecimiento denotado por un predicado. Los eventos pueden ser acciones, acontecimientos llevados a cabo voluntariamente por un sujeto agente (*los antropólogos forenses delimitaron el predio*); procesos, acontecimientos desencadenados espontáneamente (l*os árboles están floreciendo prematuramente por las altas temperaturas*) o acontecimientos causados por una fuerza externa al proceso (*Se supo que los fuertes vientos derrumbaron varios techos*); estados, situaciones que se mantienen a lo largo de un período o son permanentes (*El tránsito está detenido a causa de los cortes de ruta*).

Si bien los eventos están, en la mayoría de los casos, indicados por formas verbales, también existen nombres que designan eventos. Un nombre eventivo no designa entidades (físicas o abstractas), sino acontecimientos o sucesos como es el caso de a*ccidente, batalla, cena, eclipse, desfile, muerte, nacimiento, tempestad,* entre muchos otros.

Mientras que la morfología verbal es un poderoso indicio para la detección de eventos designados por verbos, los sustantivos que son eventivos no difieren en su morfología de aquellos que no

1 de mayo 2012 • Volumen 13 Número 5 • ISSN: 1067-60710

lo son y, por lo tanto, presentan mayor dificultad para un reconocimiento automático. Se suma a esta dificultad la ambigüedad que presentan muchos de estos nombres entre la interpretación eventiva y la de objeto: *El concierto empieza a las ocho/La orquesta interpretó el concierto en si menor para violonchelo.* Existen, sin embargo, una serie de indicios sintácticos que ayudan a reconocer este tipo de nombres: co-ocurrencia con verbos como tener lugar o presenciar; con verbos o expresiones que indican duración o fase aspectual como *empezar, comenzar, concluir, terminar, durar.*

Dijimos que un aspecto central en la comprensión computacional de un texto es la detección de los eventos referidos en él. Ahora bien, veamos el siguiente ejemplo en el cual hemos marcado con negrita los eventos y hemos subrayado uno de ellos que tiene naturaleza aspectual:

Esto dificulta *aún más el* diálogo *con el gobierno uruguayo quien* confirmó *ayer a través de la cancillería que no se* negociará *mientras* permanezca *algún* corte.

Debemos notar que mientras algunos eventos se presentan como ocurridos (*confirmó, dificulta, corte*) otros son dudosos (*diálogo*) y finalmente la eventual negociación (*negociará*) se presenta como futura y con polaridad negativa. Esto significa que no basta con encontrar términos que refieran a algún tipo de evento para inferir que dicho evento ocurrió o está ocurriendo. Es necesario, además, interpretar estos términos en sus contextos, donde pueden verse afectados por elementos de polaridad negativa, o por operadores modales, o por predicados que afectan su valor de veracidad, y combinaciones de todos ellos. La propiedad de los eventos de haber o no ocurrido o de estar ocurriendo no es entonces un dato evidente. A esta propiedad le damos el nombre de factividad (Wonsever et al. 2009).

Téngase en cuenta que, si bien la factividad está asociada con el tiempo, la modalidad y la polaridad, esta asociación no es automática. Así, eventos con los mismos valores para estos tres aspectos pueden tener diferentes comportamientos con respecto a la factividad. Las implicaciones de los predicados a los que están subordinados los eventos influyen en el valor de factividad (se subraya el predicado que subordina el evento y se indica con negrita el evento):

> *Celebro que* ***lleguen*** *mañana* [el hablante da por descontado que llegan] / *Dudo que* ***lleguen*** *mañana* [es posible que lleguen o no]; *Logró* ***cerrar*** *la puerta* [es un evento que tuvo lugar] / *Olvidó cerrar la puerta* [es un evento que no tuvo lugar]; *No dudó en* ***solicitar*** *el puesto de trabajo* [sí ocurrió] / *No quiso* ***solicitar*** *el puesto de trabajo* [no ocurrió]. También influye la información del contexto: *La* ***reunión*** *estaba planificada para las 9 y el avión llegó con retraso* [la reunión no se realizó] / *La* ***reunión*** *estaba planificada para las 9 y todos llegaron en hora* [sí se realizó].

¿De qué forma nos propusimos capturar la información con respecto al eje Eventos/Factividad?

Diseñamos un esquema de anotación, constituido por cuatro elementos. Para cada elemento

definimos una serie de atributos con diferentes valores que expresan las propiedades relevantes que caracterizan a estos elementos. Los elementos son:

- **Evento**: con esta etiqueta se anotan los eventos y se le adjudican valores a los atributos c*lase, categoría gramatical, polaridad, modalidad, factividad* (entre otros).

- **Índice**: se anota todo elemento del texto, cualquiera sea su categoría gramatical, que se considere relevante a los efectos de constituir una señal que contribuya a determinar ciertos rasgos del evento o su existencia misma.

- **Vínculo aspectual**: se anota la relación existente entre un evento de naturaleza aspectual (por ejemplo un auxiliar de fase) y el evento sobre el cual tiene alcance (*empezó* a *llover*).

- **Vínculo de subordinación:** Se anota la relación existente entre un evento y otro u otros que están subordinados a él (*intentaron robar* un banco).

Se utilizaron estos esquemas para realizar la anotación manual de un corpus. A tales efectos se elaboró una guía de anotación detallada y ampliamente ejemplificada (Wonsever et al. 2008) para orientar la tarea de anotación. Sobre este corpus anotado, se aplicaron técnicas de aprendizaje automático para generar un analizador del discurso.

Como primera experiencia de explotación del corpus anotado, desarrollamos un sistema basado en técnicas estadísticas para el reconocimiento automático de eventos (Wonsever *et al.* 2012). El sistema solamente determina los segmentos del texto correspondientes a eventos, tarea que, para el caso particular de los nombres, no es nada trivial. En el futuro, trabajaremos en la determinación automática de los valores de los atributos del evento y también en la generación de los vínculos. Los resultados son alentadores, habiéndose obtenido en el mejor caso un 80% de medida F . Este número mejora mucho (90%) si consideramos sólo los eventos verbales; la mejor medida F que obtuvimos para eventos nominales es del 56.6%. Se está anotando un mayor volumen de texto que se utilizará para hacer nuevos experimentos, así como para realizar un aprendizaje independiente de la factividad.

2.2 El módulo Enunciación

Es muy habitual, especialmente en textos periodísticos, que, además de la voz del autor, aparezcan otras voces, que se revelan bajo la forma del discurso reproducido. Poder asociar un segmento de texto con el correspondiente enunciador es, por lo tanto, una de las tareas necesarias para el procesamiento de los textos.

1 de mayo 2012 • Volumen 13 Número 5 • ISSN: 1067-60710

Veamos un ejemplo de discurso reproducido en el cual se subraya el enunciador. Se marca en negrita el predicado introductor y se encierra entre corchetes el discurso reproducido:

El investigador de la Politécnica **afirma** *[que el principal problema de este sistema es conseguir que sea fácil de usar].*

Ahora bien, no siempre aparece en forma explícita el segmento de discurso reproducido, sin embargo podemos inferirlo a partir de la ocurrencia de ciertos predicados llamados de valoración o aceptación:

El Pri acepta participar en el debate / El gobierno rechazó la propuesta

Es necesario señalar que los predicados que introducen otras voces en un texto no siempre están constituidos por verbos, sino que también pueden ser nombres (***la declaración*** *del ministro*, el **rechazo** *de la propuesta por parte del gobierno*) o incluso preposiciones (***según*** *el ministro, [no habrá aumento en el precio del combustible] /* ***de acuerdo con*** *el gobierno, [la inflación está dominada]).*

Con el objetivo de reconocer en forma automática los diferentes enunciadores que aparecen en un texto, los segmentos de discurso reproducido (mensaje) asociados a cada uno de ellos e incluso el asunto del que se habla, desarrollamos un sistema basado en reglas contextuales (Wonsever *et al*. 2001), cuyos resultados fueron mejorados aplicando un módulo estadístico (Rosá 2011). Este sistema alcanza un valor de medida F de 83% para la fuente y 61% para el mensaje. Estos valores mejoran, especialmente para el reconocimiento del mensaje, si se consideran correctos los elementos reconocidos en forma parcial, alcanzando un valor de 89%.

Además de los sistemas para el reconocimiento del discurso reproducido y sus componentes, se generó un conjunto de recursos importantes para el español: un repertorio de predicados de opinión con información sintáctico-semántica asociada; un corpus de 13.000 palabras con anotaciones sobre los componentes predicado, fuente, asunto y mensaje, y un corpus de 40.000 palabras anotado con predicados y fuentes.

2.3 El módulo Sintaxis

Este módulo cuenta con un sistema de segmentación en proposiciones (o cláusulas) y de cálculo de la estructura de una oración en términos de las proposiciones que contiene: Clatex (Caviglia et al. 2003).

La noción de proposición utilizada es la de espacio (segmento) textual donde coocurren los predicados con sus argumentos y modificadores. A su vez, dada la recursividad de las estructuras sintácticas del lenguaje, cualquiera de estos argumentos o modificadores puede ser a su vez una

proposición.

El conjunto de proposiciones recuperadas y sus relaciones estructurales en la oración, constituyen un punto de partida interesante para un analizador sintáctico, ya que este analizador no tendría que vérselas con las oraciones generalmente muy largas de los textos escritos reales, sino con las unidades proposicionales de menor tamaño que Clatex proporciona. Aunque este sistema fue desarrollado hace ya unos cuantos años, el tema de ofrecer buenas soluciones para el análisis sintáctico sigue teniendo plena vigencia.

El sistema realizado es de tipo procesamiento simbólico, aunque utiliza los resultados de un etiquetador probabilístico. Se basa en la descripción mediante reglas contextuales (tipo de reglas ya mencionado en 2.2) de los elementos centrales para la determinación de las proposiciones (verbos, conjunciones, signos de puntuación) y su estructuración en la oración.

La noción de proposición que se utiliza en este trabajo es sintáctica y "concreta": se considera proposición a un segmento de texto que contiene un verbo, sus argumentos y modificadores. Si bien la idea de proposición está vinculada a una noción más amplia de predicación, es claro que la predicación no se realiza solamente por medio de expresiones que giran en torno a un verbo conjugado, así como no siempre son predicativas este tipo de expresiones. Es por esto que decimos que utilizamos una noción sintáctica de proposición. El término "concreta" se refiere a que la noción gramatical abstracta de proposición se identifica con una realización material en términos de segmentos de texto.

Veamos el resultado del análisis que realiza nuestro sistema sobre el siguiente texto de entrada:

Cuando se editó hace muy poco un disco compacto de Ruben Rada en Buenos Aires, con materiales grabados en 1987, no se consultó a Rada ni a sus colaboradores ni se les pagó por la edición, que por otra parte fue muy grande puesto que se tuvo que vender a precio muy bajo, compitiendo en el mercado con su disco recién editado en Estados Unidos.

[enunciado

[prop1

[prop2

[prop3 Cuando se editó hace muy poco un disco compacto de Rubén Rada en Buenos Aires, con materiales grabados en 1987

/prop3] ,

no se consultó a Rada ni a sus colaboradores

/prop2],

ni

[prop4 se les pagó por la edición ,

1 de mayo 2012 • Volumen 13 Número 5 • ISSN: 1067-60710

[prop5 que por otra parte fue muy grande

[prop6 puesto que se tuvo que vender a precio muy bajo, compitiendo en el mercado con su disco recién editado en Estados Unidos

/prop6]

/prop5]

/prop4]

/prop1]

/enunciado]

El texto en negrita **prop** corresponde a las marcas de proposición generadas por el sistema. Hay un par de marcas (apertura y cierre) por cada proposición.

En la oración anterior se reconocieron 6 proposiciones. Las marcas **enunciado** señalan comienzo y fin de la oración de texto. Consideramos que la puntuación es una indicación explícita de la estructuración en unidades en textos escritos.

La validación del sistema se realizó a partir del análisis de un conjunto de artículos periodísticos, no incluidos en el corpus utilizado para la inferencia de reglas. Estos textos contienen 176 oraciones, con un promedio de 40 palabras por oración. Se obtuvieron tasas de precisión y de recuperación del 91%. Restringiendo los cálculos a oraciones consideradas “complejas” (más de 3 proposiciones), estos resultados empeoran y la proporción de proposiciones, correctamente recuperadas sobre el total de proposiciones, desciende al 87%.

3. Interacción entre los módulos

Los distintos módulos, tratados de manera independiente, se vinculan necesariamente para dar cuenta de la interpretación del texto de manera apropiada.

El módulo sintáctico, en donde operan Clatex y Freeling, analizador morfosintáctico de uso libre, ofrece un servicio de base a los otros módulos, identificando proposiciones subordinadas y clasificando las clases de palabras con los rasgos flexivos asociados.

Para el módulo Eventos-Factividad tiene especial relevancia tanto la información referida al modo y al tiempo verbal, a los efectos de calcular la factividad, como la delimitación de las proposiciones para determinar el alcance de ciertos predicados. Asimismo, para el módulo Enunciación, la segmentación de Clatex resulta relevante para establecer la extensión del segmento que contiene el discurso reproducido.

La factividad representa la perspectiva del enunciador con respecto al evento. Como vimos, el enunciador no coincide necesariamente con el autor del texto y es el módulo de la Enunciación el que nos proporciona la fuente con la que debe vincularse cada evento.

4. Desarrollo actual y perspectivas

En el momento actual estamos aumentando el volumen del corpus anotado a los efectos de realizar nuevos experimentos de aprendizaje automático, que impliquen el reconocimiento de algunos atributos del elemento evento.

Por otra parte, estamos trabajando en la generación de recursos léxicos que incluyen, por ejemplo, un repertorio de predicados implicativos, los cuales, combinados con otros elementos, permiten calcular el valor de factividad de un evento.

El módulo Temporalidad involucra, por un lado, el reconocimiento de las expresiones temporales en los textos y, por otro, el reconocimiento de las relaciones temporales entre los eventos y entre los eventos y los intervalos denotados por las expresiones temporales. Para lograr este objetivo se ha elaborado un modelo descriptivo y un esquema de anotación de expresiones y marcas temporales (Wonsever et al. 2011). Se está trabajando en la anotación manual del corpus para utilizarlo como conjunto de entrenamiento en el aprendizaje automático.

Referencias

Caviglia, S., J. Couto, A. Rosá, D. Wonsever (2003): "Reconocimiento de indicadores de inicio de proposición en español", Actas del VIII Simposio de Lingüística Aplicada, Santiago de Cuba.

Rosá, A. (2011): "Identificación de opiniones de diferentes fuentes en textos en español", Tesis Doctoral, Universidad de la República (Uruguay) y Université Paris Ouest Nanterre La Défense (Francia).

Wonsever D., J-L. Minel (2001): "Contextual Rules for Text Analysis", Lecture Notes in Computer Science 2004, february.

Wonsever D., M. Malcuori, A. Rosá (2008): "SIBILA: Esquema de anotación de eventos", Reporte técnico 08–11, ISSN: 0797–6410, Biblioteca InCo PEDECIBA.

Wonsever D., M. Malcuori, A. Rosá (2009): "Factividad de los eventos referidos en textos", Reporte técnico 09–12, ISSN: 0797–6410, Biblioteca InCo PEDECIBA.

Wonsever D., M. Malcuori, M. Etcheverry (2011): "Guía de anotación de expresiones y marcas temporales. Proyecto Temantex", Reporte técnico 11–15, ISSN: 0797–6410, Biblioteca InCo PEDECIBA.

Wonsever D., A. Rosá, M. Malcuori, G. Moncecchi, A. Descoins (2012): "Event Annotation Schemes and Event Recognition in Spanish Texts", en A. Gelbukh (ed.) Computational Linguistics and Intelligent Text Processing, LNCS 7182, Springer. New Delhi.

1 de mayo 2012 • Volumen 13 Número 5 • ISSN: 1067-60710