Planificación posicional en el diseño de sistemas versátiles de generación de lenguaje natural

Planning with Positional Language Models to produce versatile Natural Language Generation systems

Marta Vicente
Universidad de Alicante
Departamento de Lenguajes y Sistemas Informáticos
mvicente@dlsi.ua.es

Resumen: La Generación de Lenguaje Natural es la disciplina que permite conferir cierta forma textual a un conjunto de datos, persiguiendo un determinado objetivo comunicativo. El actual escenario tecnológico reclama sistemas de generación flexibles, adaptables a diferentes casos de uso, por lo que la incorporación dinámica del contexto debe formar parte del sistema mismo. La propuesta que aquí se expone busca introducir información semántica y pragmática en la parte del proceso que selecciona y estructura los mensajes esperados en la salida. Para ello, se estudian técnicas estadísticas que modelizan la distribución y posición de ciertos elementos semánticamente relevantes con el objetivo de reconocer patrones y relaciones generalizables.
Palabras clave: Generación de Lenguaje Natural, macroplanificación, sistemas híbridos, semántica, pragmática

Abstract: The automatic generation of natural language is the area within computational linguistics responsible for providing certain textual shape to some information that has been selected and processed to achieve a communicative goal. On the one hand, the concern is to determine what is to be communicated while on the other, the system will decide how to say it. The current technological scenario demands flexible and adaptable systems. We propose a statistical approach that leverages on the distribution of relevant elements to detect patterns and promote domain and application independence, with particular emphasis on the early stages of the process.
Keywords: Natural Language Generation, macroplanning, semantics, pragmatics

1 Motivación

En términos generales, denominamos Tecnologías del Lenguaje Humano (TLH) a aquel conjunto de técnicas, métodos y aplicaciones que, desde una perspectiva computacional, trabajan el lenguaje natural ya para extraer conocimiento o para generarlo. Aunque es cierto que ambas perspectivas, en mayor o menor medida, se expresan simultáneamente en multitud de áreas (*question answering*, diálogos, traducción automática), desde nuestro planteamiento, la Generación de Lenguaje Natural (GLN) constituye una disciplina susceptible de ser caracterizada y analizada *per se*.

De este modo, la GLN se definiría como la disciplina responsable de presentar adecuadamente una determinada información que ha sido procesada bajo la premisa de conseguir un objetivo comunicativo específico (por ejemplo, informar o resumir). Para alcanzar su propósito, un sistema de GLN tiene que acometer dos tipos de acciones: determinar qué se ha de comunicar y decidir cómo construir la salida requerida (Reiter y Dale, 2000). La presente propuesta de investigación se centra en la primera, conocida como macroplanificación, en la que se seleccionan los contenidos y se establece una estructura para los mismos. Esta información se representa en el *plan de documento* que ha de servir como guía en lo que resta del proceso.

Tradicionalmente, el diseño de un sistema

de GLN se ha llevado a cabo determinado por el ámbito, el dominio o el género para el que se desarrollaba. Este trabajo se centra en estudiar y producir tecnologías capaces de realizar las tareas asociadas a la macroplanificación de forma flexible e independiente. Para ello, se investiga la incorporación de aspectos semánticos y pragmáticos al proceso, que permitan que el sistema resultante se pueda adaptar a diferentes escenarios de uso. Se impulsa, por tanto, el uso de técnicas estadísticas que ya en otras áreas del lenguaje han demostrado su efectividad en términos de adaptación, ya sea al género, al dominio o al idioma en que se requiere el texto.

2 Antecedentes y trabajos relacionados

En general, un sistema de GLN puede abordarse desde dos aproximaciones, empleando técnicas estadísticas o basadas en conocimiento. El planteamiento estadístico establece una relación entre el texto generado y cierta probabilidad asociada con los elementos que lo constituyen (Kondadadi, Howald, y Schilder, 2013; Barros y Lloret, 2015). Por otro lado, los sistemas basados en conocimiento obtienen la información que necesitan de lexicones o tesauros, o emplean plantillas o conjuntos de reglas asociados generalmente a dominios específicos(Dannélls et al., 2012; Bouayad-Agha, Casamayor, y Wanner, 2011).

Los sistemas estadísticos son más flexibles respecto al dominio, aunque los basados en conocimiento son más permeables a cuestiones semánticas y pueden incorporar un conocimiento contextual y lingüístico del que carecen los primeros. En este momento, estudiamos la aplicación de un tipo de modelo estadístico que permite introducir información relativa a la distribución de los elementos en el documento, incorporando cierta información del contexto de los mismos en el modelo resultante.

3 Descripción de la investigación

Partiendo de este panorama general, esta tesis doctoral plantea como objetivo la investigación y desarrollo de un sistema de *GLN* flexible, centrándose sobre todo en la tarea de selección y en la elaboración del plan del documento como estructura contenedora del conocimiento pragmático y semántico imprescindible para que el sistema proporcione salidas adaptadas a las circunstancias en que se requiere el discurso.

3.1 Hipótesis

La hipótesis de partida, por tanto, es que la incorporación de información semántica y pragmática tanto en el diseño, como en la investigación y el desarrollo de un sistema de *GLN*, y en concreto en este trabajo, a la fase de selección y estructuración del contenido, redunda en la consecución de sistemas más flexibles y adaptables, evitando las dificultades que hasta ahora se manifestaban a través de dependencias respecto al dominio o a la aplicación. Flexibilidad referida a:

- las bases de conocimiento y fuentes de información relevante: datos sin tratar o estructurados, bases de datos, la web, etc.
- Las técnicas para determinar la estructura y el contenido, planteamientos híbridos que aprovechen las ventajas y superen los inconvenientes de las aproximaciones previas.
- La incorporación de información semántica y pragmática que redunde en la adaptabilidad del sistema a diferentes escenarios, considerando las intenciones del usuario y el propósito del sistema.

Para poder abordar tal tarea, se requiere un estudio de los fundamentos lingüísticos que vertebren esa dimensión del proceso de investigación. De este modo, también se realiza el análisis de aquellas teorías que se centran tanto en el aspecto pragmático del lenguaje como en el funcional, en tanto condicionantes de la forma y contenido del discurso. Sería el caso de la teoría de la estructura retórica (Mann y Thompson, 1988) o la lingüística sistémico funcional (Halliday y Matthiessen, 2013).

Por otro lado, se mantiene un seguimiento de las técnicas aplicadas tanto en el ámbito de la generación como en otras áreas del procesamiento de lenguaje natural, como puede ser la incorporación de diferentes tipos de modelos de lenguaje (Manning y Schütze, 1999; Lv y Zhai, 2009) o las omnipresentes redes neuronales (Goldberg, 2016), con el doble propósito de introducirlas en nuestro esquema y de realizar una comparación del comportamiento de tales aproximaciones.

4 Metodología propuesta

La metodología que estamos siguiendo considera la contextualización de la investigación (los antecedentes, los enfoques actuales relativos a cada aspecto revisado) así como el diseño de experimentos y los métodos de evaluación que nos permitan refinar técnicas y precisar las características del sistema. Algunos objetivos generales:

- estudio y sistematización de los diferentes enfoques que actualmente se aplican en *GLN* y concretamente, en macroplanificación.
- Estudio y adaptación de teorías lingüísticas que vertebren el desarrollo del sistema sobre fundamentos semánticos y pragmáticos estables.
- Estudio y adaptación de técnicas híbridas (estadísticas/conocimiento) así como de técnicas relevantes en el área de inteligencia artificial y aprendizaje automático.

5 Modelos de lenguaje posicionales (MLPs).

Actualmente, estamos realizando una serie de experimentos centrados en el estudio y aplicación de un tipo de modelo de lenguaje que, frente a aquellos modelos que consideran el texto como una bolsa de palabras, permite recuperar información relativa a los elementos relevantes y su distribución en el texto. Han sido utilizados previamente en otras tareas como la búsqueda y recuperación de información (Lv y Zhai, 2009) o la generación de resúmenes extractivos (Liu et al., 2015).

Los MLPs se calculan empleando una función de propagación, de modo que podemos atribuir un valor P a un elemento w que aparece en una determinada posición i, en función del resto de ocurrencias del mismo elemento en el texto y su distancia respecto a i, tal y como se muestra en la Ecuación 1:

$$P(w \mid i) = \frac{\sum_{j=1}^{|D|} c(w,j) \times f(i,j)}{\sum_{w' \in V} \sum_{j=1}^{|D|} c(w',j) \times f(i,j)} \quad (1)$$

donde *c(w,j)* indica la presencia del término w en la posición j, $|D|$ se refiere a la longitud del documento, V se refiere a el vocabulario y *f(i,j)* es la función de propagación.

Una vez calculados los modelos para cada posición i, es posible recuperar los elementos más relevantes del texto, los que mayor puntuación hayan obtenido, según diferentes segmentaciones. Por el momento, nuestra referencia es la longitud de las oraciones originales, pero podríamos considerar párrafos, conjuntos de oraciones relacionadas con un tema, etc.

Con los elementos seleccionados se construye el *plan de documento*. Los elementos proceden del vocabulario previamente determinado. Este vocabulario se deriva del contenido original del texto. Puede ser el conjunto de lemas, de synsets que lo constituyen o estructuras más elaboradas como eventos, relaciones, agentes asociados a los roles semánticos, etc. En cada línea del plan de documento incluiremos los elementos más importantes asociados a esa área del texto.

La Figura 1 representa el procesamiento que conduce a la obtención de un plan de documento a partir de un texto de entrada. En la presente implementación la *matriz de importancia* (Mi) está formada por tantas filas como elementos tiene el vocabulario V y tantas columnas como posiciones el texto $|D|$. La *matriz de puntuaciones* contiene los valores $P(w \mid i)$ para cada una de las palabras del vocabulario, entonces, habiendo considerado una función de distancia para computar su resultado. Los elementos con mayor *puntuación* componen el plan de documento final.

5.1 Experimentos

En el momento actual se está llevando a cabo una aplicación de tales técnicas en el ámbito de los cuentos para niños (Vicente, Barros, y Lloret, 2017). Trabajamos con un corpus en inglés (779 cuentos) que procede de 3 fuentes: el corpus de Lobo y Matos (Lobo y De Matos, 2010), cuentos de Andersen (*Hans Christian Andersen: Fairy Tales and Stories*[1]) y un conjunto extraido automáticamente del sitio *Bedtime stories*[2].

En cuanto al vocabulario, los elementos seleccionados son, por un lado, los lemas junto a la categoría morfológica y, por otro, los synsets de WordNet (Fellbaum, 1998) para los términos de los cuentos, extraídos empleando Freeling (Padró y Stanilovsky, 2012), dando lugar a dos posibles experimentos.

Respecto a la función de propagación, se ha optado por el kernel Gaussiano siguiendo trabajos previos (Liu et al., 2015; Lv y Zhai, 2009).

[1] http://hca.gilead.org.il/

[2] https://freestoriesforkids.com/

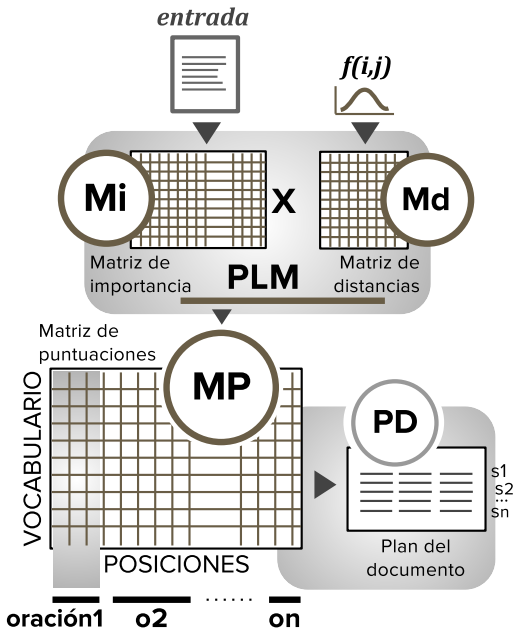

Figura 1: Secuencia de etapas de procesamiento por la que, a partir de una entrada en forma de texto, se elabora un plan de documento tomando como base modelos de lenguaje posicionales.

Dada la segmentación que hemos elegido, por cada línea del documento original dispondremos de una línea en el plan de documento.

Tanto si estamos en la configuración con lemas como en la de synsets, se proporcionarán varios elementos de las diferentes categorías morfológicas que requiera el módulo de realización en cada caso (verbos, nombres, adjetivos, etc).

5.2 Evaluación

En cuanto a la evaluación, existe una investigación permanente relativa a los métodos adecuados en el campo de la *GLN*, pues no se dispone de un corpus de referencia con el que cotejar la salida. En ese sentido, los experimentos y etapas cubiertas en el transcurso de la investigación son evaluados tanto mediante estrategias extrínsecas como intrínsecas.

Por un lado, se considera su impacto sobre el sistema de GLN, pasándoselo al módulo de realización como entrada. en este caso, se realiza una evaluación manual de los resultados, considerando la fluidez y legibilidad de la salida. Por otro lado, se miden diferentes indicadores como la variedad en el vocabulario del texto resultante. También se emplean métricas que, aunque se suelen emplear en otras tareas ajenas a la *GLN*, pueden dar cierta luz sobre el tipo de resultados que obtenemos, como ROUGE (Lin, 2004), empleada en resúmenes, o TER (Snover et al., 2006), empleada en traducción automática.

6 Líneas abiertas y cuestiones específicas

Hemos introducido en la Sección 5 un tipo de modelo de lenguaje así como los primeros experimentos realizados empleándolo para obtener planes de documento, esto es, estructuras de contenido que permitan a un módulo de realización construir un texto nuevo. Son muchas las posibilidades que presenta tal planteamiento y, con el análisis de los resultados y la incorporación de nuevas herramientas de análisis, estamos ampliando su capacidad y refinando su comportamiento.

Por ejemplo, las funciones de propagación asociadas al modelo incluyen un parámetro σ que, en este caso, determina el alcance semántico de un término en una posición. Siguiendo los trabajos mencionados, ese valor se fijó en un primer momento a 25, pero estamos analizando otros valores, así como la posibilidad de que se haga depender al mismo del contexto (longitud de la oración que lo contiene, relevancia del término considerado, etc). También se prueban otros tipos de funciones de propagación, como el kernel circular o el triangular. La elaboración de las características del modelo se puede extender no solo a la función de propagación, sino también al tipo de segmentación (oraciones, sintagmas, secciones por temáticas, ...), al tipo de elementos que componen el vocabulario (verbos, sustantivos, entidades, eventos, ...), al número y disposición de los términos incluidos en el plan de documento, etc. Otra de las líneas en las que se trabaja es la normalización o transformación de los resultados obtenidos para cada documento del corpus de manera que podamos aplicar técnicas de aprendizaje para detectar los esquemas y patrones asociados a los cuentos. Esto nos permitiría obtener una descripción mas generalizable de los mismos.

Considerando, por tanto, el marco presentado, surgen algunas cuestiones cuya puesta en común resultaría enriquecedora: ¿qué tipo de características podemos incorporar para conseguir mejoras desde el punto de vista pragmático, qué tipo de elementos debería contener el vocabulario, que herramientas nos permitirían extraerlos? Bajo el mismo pris-

ma, ¿qué técnicas de aprendizaje son apropiadas para una investigación que pretende nutrirse de textos de muy diversa índole y estructura, cómo podrían presentarse las características para tales métodos? ¿Qué técnicas de evaluación serían convenientes para probar su comportamiento y validez? ¿Qué recursos se están empleando en otros idiomas en estudios similares? Esperamos que las ideas que tales preguntas susciten reviertan en una mejora cualitativa tanto de la investigación como de los resultados futuros.

Agradecimientos

Esta investigación ha sido financiada por la Generalitat Valenciana mediante el contrato ACIF/2016/501 y el proyecto "DIIM2.0: Desarrollo de técnicas Inteligentes e Interactivas de Minería y generación de información sobre la web 2.0" (PROMETEOII/2014/001), y por el Gobierno de España (MINECO) a través del proyecto "RESCATA: Representación canónica y transformaciones de los textos aplicado a las tecnologías del lenguaje humano" (TIN2015-65100-R).

Bibliografía

Barros, C. y E. Lloret. 2015. Input seed features for guiding the generation process: A statistical approach for spanish. *Proceedings of the 13th European Workshop on Natural Language Generation*, página 9.

Bouayad-Agha, N., G. Casamayor, y L. Wanner. 2011. Content selection from an ontology-based knowledge base for the generation of football summaries. En *Proceedings of the 13th European Workshop on Natural Language Generation*, páginas 72–81. ACL.

Dannélls, D., L. Carlson, K. Ji, J. Saludes, K. Kaljurand, M. Damova, A. Kiryakov, M. Grinberg, M. K. Bergman, F. Giasson, y others. 2012. Multilingual text generation from structured formal representations. *University of Gothenburg*, 7427.

Fellbaum, C. 1998. *WordNet: An Electronic Lexical Database*. MIT Press.

Goldberg, Y. 2016. A primer on neural network models for natural language processing. *J. Artif. Intell. Res.(JAIR)*, 57:345–420.

Halliday, M. A. y C. M. Matthiessen. 2013. *Halliday's introduction to functional grammar*. Routledge.

Kondadadi, R., B. Howald, y F. Schilder. 2013. A statistical nlg framework for aggregated planning and realization. En *ACL (1)*, páginas 1406–1415.

Lin, C.-Y. 2004. Rouge: A package for automatic evaluation of summaries. En *Text Summarization Branches Out: Proceedings*, páginas 74–81. ACL.

Liu, S.-H., K.-Y. Chen, B. Chen, H.-M. Wang, H.-C. Yen, y W.-L. Hsu. 2015. Positional language modeling for extractive broadcast news speech summarization. En *INTERSPEECH*, páginas 2729–2733.

Lobo, P. V. y D. M. De Matos. 2010. Fairy tale corpus organization using latent semantic mapping and an item-to-item top-n recommendation algorithm. En *Language Resources and Evaluation Conference - 2010*.

Lv, Y. y C. Zhai. 2009. Positional language models for information retrieval. En *Proceedings of the 32Nd International ACM SIGIR*, páginas 299–306. ACM.

Mann, W. C. y S. A. Thompson. 1988. Rhetorical structure theory: Toward a functional theory of text organization. *Text-Interdisciplinary Journal for the Study of Discourse*, 8(3):243–281.

Manning, C. D. y H. Schütze. 1999. *Foundations of statistical natural language processing*, volumen 999. MIT Press.

Padró, L. y E. Stanilovsky. 2012. Freeling 3.0: Towards wider multilinguality. En *Proceedings of LREC 2012*. European Language Resources Association.

Reiter, E. y R. Dale. 2000. *Building Natural Language Generation Systems*. Cambridge University Press.

Snover, M., B. Dorr, R. Schwartz, L. Micciulla, y J. Makhoul. 2006. A study of translation edit rate with targeted human annotation. En *Proceedings of AMTA*, páginas 223–231.

Vicente, M., C. Barros, y E. Lloret, 2017. *A Study on Flexibility in Natural Language Generation Through a Statistical Approach to Story Generation*, páginas 492–498. Springer International Publishing.