Estudio de un enfoque híbrido para la Generación del Lenguaje Natural

Study of a hybrid approach for Natural Language Generation

Cristina Barros
Departamento de Lenguajes y Sistemas Informáticos
Universidad de Alicante
Carretera de San Vicente del Raspeig s/n - 03690 Alicante (España)
cbarros@dlsi.ua.es

Resumen: Este proyecto de tesis plantea una aproximación híbrida para la generación del lenguaje natural, la cual permitirá mejorar la calidad del texto producido, favoreciendo la independencia del dominio, del género textual y de la aplicación final donde se utilice. Con el fin de lograr este objetivo, se ha implementado un enfoque flexible de generación centrado en la fase de realización, el cual, apoyándose en conocimientos estadísticos y en lexicones, permite generar textos para diferentes dominios e idiomas guiados por la entrada.
Palabras clave: Generación del lenguaje natural, *"característica semilla"*, modelos de lenguaje factorizados, realización, aproximación híbrida

Abstract: The main objective of this thesis is to propose a hybrid natural language generation approach which will improve the quality of the produced text, encouraging the independence of domain, text type and application. In order to achieve this objective, we present a flexible natural language generation approach focused on the surface realisation stage, which, based on statistical knowledge and lexicons, allows the generation of text for different domains and languages guided by the system input.
Keywords: Natural language generation, seed feature, factored language models, surface realisation, hybrid approach

1 Motivación

Dada la necesidad existente para facilitar la comunicación y la interacción hombre-máquina (Jacko, 2012), las Tecnologías del Lenguaje Humano, encargadas de procesar el lenguaje humano de forma automática, tienen un papel clave. De entre todas las subdisciplinas pertenecientes a las *TLH*, el área de la Generación del Lenguaje Natural *(GLN)* es capaz de producir lenguaje a partir de entradas no lingüísticas.

Gracias a las características que ofrece el área de la *GLN*, esta puede emplearse en distintos ámbitos, como en la meteorología (Goldberg, Driedger, and Kittredge, 1994; Reiter et al., 2005), donde a partir de datos numéricos procedentes de sensores y de sistemas de simulación, que representan distintas magnitudes como la temperatura, la velocidad del viento o el nivel de precipitaciones de un determinado lugar, se puede generar un informe explicativo. Asimismo, se han empleado este tipo de ténicas en medicina (Gatt et al., 2009; Acharya et al., 2016), pudiendo generar texto, a partir de datos obtenidos mediante sensores, que se adecuen a distintos registros dependiendo del perfil del usuario.

Además, también se han diseñado sistemas de *GLN* como herramienta de ayuda para la comunicación de personas con algún tipo de discapacidad o problemas de comprensión lectora (Reiter et al., 2009; Ferres et al., 2006), así como también pueden incorporar ténicas para que personas cuasi analfabetas puedan leer (Williams and Reiter, 2008).

2 Antecedentes y trabajos relacionados

La tarea de la *GLN*, a grandes rasgos, consiste en producir de forma automática estructuras correctas del lenguaje natural a partir de una representación de la información (Cole et al., 1997), ya sea en texto o en forma de algún tipo de dato, permitiendo así que se proporcione a los usuarios nueva información inferida.

Esta tarea se ha dividido comúnmente en

varias etapas diferenciadas: la macro planificación, la micro planificación y la realización (Reiter and Dale, 2000), siendo el objetivo de estas determinar la información contenida en el nuevo texto a generar (macro planificación) y cómo queremos representar dicha información en un nuevo texto (micro planificación y realización).

Tradicionalmente una de las limitaciones de los sistemas de *GLN* es que se han diseñado para dominios muy concretos y para un fin determinado, siendo el desarrollo de enfoques de dominio abierto y flexibles un reto para la comunidad investigadora.

Actualmente, uno de los enfoques más recientes para abrodar la tarea de la *GLN* en los últimos años es la generación empleando técnicas estadísticas (Bohnet, Mille, and Wanner, 2011; Wan et al., 2009; Lemon, Janarthanam, and Rieser, 2012), cuya idea subyaciente se basa en analizar y calcular la probabilidad de que ciertas palabras aparezcan juntas. A partir de este tipo de probabilidades se puede realizar un estudio de la formación de una frase a partir de un conjunto de palabras iniciales. Junto a este tipo de enfoques estadísticos, existen otros enfoques basados en el uso de conocimiento, los cuales recurren a teorías lingüísticas, como puede ser la Teoría de la Estructura Retórica del discurso (Mann and Thompson, 1988) o la Teoría sentido-texto de Mel'čuk (Žolkovskij and Mel'čuk, 1965), para generar un texto dado.

3 Propuesta de investigación

La hipótesis de partida de esta investigación es que la aplicación de una aproximación híbrida para la *GLN* permitirá incrementar la calidad del lenguaje producido, favoreciendo su independencia del dominio, del género textual y de la aplicación final que lo utilice, siendo la implementación de un enfoque de generación híbrido el objetivo final de la tesis.

4 Metodología y experimentos

Con el objetivo de lograr una aproximación híbrida que favorezca la independencia del dominio, género textual y aplicación, se ha implementado un método flexible centrado en la fase de realización de la *GLN* cuya novedad, con respecto al estado de la cuestión radica en que en la entrada al sistema es una *"característica semilla"*. Esta *"característica semilla"* puede ser vista como un objeto abstracto (un fonema, una palabra, un sentimiento, etc.) encargado de guiar el proceso de generación con respecto al contenido del texto generado. Por tanto, guiará la generación en relación con su vocabulario o el tipo de palabras que deberá contener el nuevo texto generado, aportando así la flexibilidad necesaria al enfoque para poder adaptar con facilidad la generación de textos independientemente del dominio e idioma. En la Figura 1 se puede ver un esquema general del enfoque de generación que se describirá en las próximas líneas.

Figura 1: Diagrama del enfoque de generación propuesto en el que se generan frases flexionadas guiadas por la *"característica semilla"* de entrada.

En este enfoque se ha estudiado la aplicación de técnicas estadísticas para la generación, las cuales, en conjunción con información obtenida de diversas fuentes (e.g. lexicones, herramientas, etc.) nos permite una generación flexible. En este caso, se ha probado el método empleando modelos de lenguaje factorizados (*FLM: Factored Language Models*), que son una extensión de los modelos de lenguaje introducidos en (Bilmes and Kirchhoff, 2003), donde una palabra es vista como un vector de k factores tal que $w_t \equiv \{f_t^1, f_t^2, \ldots, f_t^K\}$. Estos factores pueden ser cualquier cosa, incuyendo lemas, etiquetas gramaticales, o cualquier otra característica léxica, sintáctica o semántica. Una vez que se selecciona un conjunto de factores, el objetivo principal de los *FLM* es crear un modelo estadístico $P(f|f_1, \ldots, f_N)$ donde la predicción de una característica f esté basada en sus N padres $\{f_1, \ldots, f_N\}$. Estos FLM se emplean para generar las oraciones, priorizando la selección de palabras que esten relacionadas con la *"característica semilla"* deseada para la generación.

Dependiendo de los factores empleados para la generación, el texto generado puede

no contener elementos flexionados, siendo la flexión automática de frases otro punto clave para lograr el objetivo marcado. En este caso, se ha implementado un módulo de flexión de palabras para diferentes idiomas, español e inglés. En el caso del inglés, la flexión se realiza con reglas escritas a mano dado que las flexiones en este idioma tienen muy pocas variantes. Sin embargo, debibo a la complejidad que entraña la flexión en lenguajes morfológicamente ricos como es en este caso el español, se ha realizado la flexión de las frases empleando técnicas de aprendizaje automático en el caso de los verbos, mientras que, para el resto de palabras se han empleado reglas escritas a mano. Específicamente, para el aprendizaje de la flexión de verbos españoles, en una primera instancia, se elaboró un conjunto de datos que contenía todas las reglas necesarias para poder realizar la flexión de todos los verbos independientemente de su conjugación y del tipo de verbo que sea (regular e irregular). Este conjunto de datos fue creado consultando la *Real Academia Española*[1] y la *Enciclopedia Libre Universal en Español*[2].

El conjunto de datos está compuesto por las siguientes características: (1) *ending*, (2) *ending stem*, (3) *penSyl*, (4) *person*, (5) *number*, (6) *tense*, (7) *mood*, (8) *suff1*, (9) *suff2*, (10) *stemC1*, (11) *stemC2*, (12) *stemC3*.

Se ha considerado que un verbo español se puede dividir en tres partes: (1) *ending* (que hace referencia a la conjugación); (2) *ending stem* (i.e. la consonante más cercana a la característica *ending*); and (3) *penSyl* (i.e. la penúltima sílaba del verbo que puede estar formada por la sílaba entera o por su vocal dominante), como se muestra en la Figura 2, siendo estas partes las que pueden variar en la flexión del verbo.

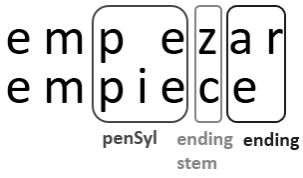

Figura 2: División del verbo empezar y su flexión para la primera persona del singular del presente de subjuntivo.

Se entrenó un conjunto de modelos individuales para cada una de las características con un valor de flexión potencial. Se usó la implementación de WEKA (Frank, Hall, and Witten, 2016) del algoritmo *Random Forest* para entrenar los modelos de las características *stemC3* y *stemC2*. Para entrenar los modelos de las características *suff1*, *suff2* y *stemC1* se empleó la implementación del algoritmo *Random Tree*. Con estos modelos entrenados se pueden predecir todas las posibles felxiones de un verbo dado su infinitivo. Para llevar a cabo esta tarea, primero se analiza el infinitivo del verbo para poder extraer las características necesarias para la flexión, y entonces se predice la flexión de cada característica usando los modelos entrenados. Finalmente, las flexiones predichas se sustituyen en el infinitivo del verbo por las características previamente identificadas, lo que conduce a una reconstrucción del infinitivo en la flexión deseada, como se muestra en la Figura 3.

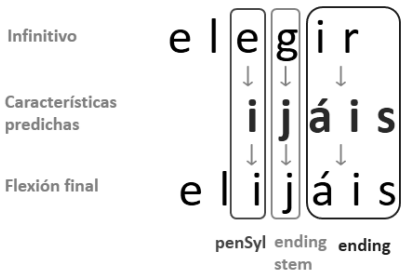

Figura 3: Reconstrucción del verbo "elegir" con las características predichas por los modelos.

Este enfoque contribuye al estado de la cuestión en lo siguiente: se presenta un método flexible capaz de generar lenguaje que es fácilmente adaptable a diferentes dominios e idiomas; se presenta un módulo de flexión eficiente, para diversos idiomas, que emplea reglas escritas a la vez que es capaz de predecir la flexión de palabras que no se adecuen a las reglas, para el caso de los verbos españoles.

4.1 Progreso de la investigación

Para validar el enfoque propuesto se han realizado experimentos con respecto a la aplicación de métodos estadísticos así como también experimentos para validar el módulo de flexión.

Con respecto al empleo de los *FLM*, se han escogido varios factores con información sintáctica y semántica (incluyendo palabras, lemas, etiquetas gramaticales (*POS tag:*

[1] http://www.rae.es/diccionario-panhispanico-de-dudas/apendices/modelos-de-conjugacion-verbal

[2] http://enciclopedia.us.es/index.php/Categoría:Verbos

Part-of-Speech tag) y synsets[3]) para entrenar varios modelos *FLM* y evaluar las frases generadas atendiendo a diferentes criterios. Se generaron un total de 20 frases por cada una de las configuraciones de factores: i) *Palabras + POS tag*, ii) *Lemas + POS tag* y iii) *Synset + POS tag*.

Para evaluar las frases generadas, realizamos una evaluación de usuario colaborativa con un total de 12 participantes como asesores. Para dicha evaluación se emplearon cuestionarios con varias preguntas empleando una escala de Likert de 5 niveles. Estas preguntas estaban relacionadas con la *coherencia* y los *errores gramaticales* contenidos en las frases generadas. El término *coherencia* se refiere al nivel de significado de las frases, siendo 1 el valor para frases con poco sentido y un 5 el valor para frases con un significado completo. Por otra parte, el término de *errores gramaticales* se refiere a la cantidad de errores gramaticales que tienen las frases generadas, siendo 1 el valor usado cuando las frases contienen un alto número de errores y 5 el valor empleado para denotar la ausencia de errores en ellas.

Factores	**Coherencia**	**Errores Gramaticales**
Palabra+POS	2,68	2,83
Lema+POS	3,08	3,00
Synset+POS	2,85	3,08

Tabla 1: Resultados de las medias de la escala de Likert de 5 niveles con respecto a la coherencia y errores gramaticales de las frases generadas estadísticamente empleando distintos factores en los FLM.

En la Tabla 1 se puede observar un resumen de las medias obtenidas para los criterios mencionados. Estos resultados muestran que el empleo de factores más abstractos y generales (los lemas y synsets en conjunción con el *POS tag*) a la hora de generar nos aporta una mayor capacidad expresiva.

Por otro lado, en el caso del módulo de flexión, debido a que la flexión de oraciones en español es más compleja, se realizó un experimento donde se generaron un total de 81 frases en español empleando la configuración de *Lema + POS tag* para el *FLM*. Las frases generadas se flexionaron, tal y como se mencionó en el apartado 4, de dos maneras distintas: i) dejando la flexión del verbo en un tiempo verbal fijo para todas las frases y ii) flexionando cada frase con un tiempo verbal aleatorio entre todos los tiempos verbales simples del español.

En este caso se volvió a realizar una evaluación de usuario colaborativa con un total de 3 participantes como asesores. Para esta evaluación se empleó el mismo tipo de cuestionarios que en el experimento de los modelos estadísticos, utilizando los mismos criterios de evaluación (*coherencia* y *errores gramaticales*) con una escala Likert de 5 niveles. Se evaluaron tanto las frases sin flexionar como las frases con los dos tipos de flexión comentados.

Tipo de Flexión	**Coherencia**	**Errores Gramaticales**
Sin flexión	2,65	2,73
Fija	3,36	3,57
Aleatoria	3,31	3,51

Tabla 2: Resultados de las medias de la escala de Likert de 5 niveles con respecto a la coherencia y errores gramaticales de las frases generadas flexionadas.

En la Tabla 2 se puede ver un resumen de los resultados obtenidos, los cuales indican una gran mejoría en la calidad y expresividad de las frases flexionadas con respecto a su variante sin flexionar.

5 Cuestiones de investigación

Siendo la *GLN* un área de interés en el Procesameinto del Lenguaje Natural, y dado que estos resultados son prometedores, las siguientes cuestiones a investigar serían: i) la investigación de métodos de evaluación automática para la *GLN* con el fin de discernir la validez del texto generado, y ii) analizar diversos métodos basados en conocimiento que nos permitan mejorar el lenguaje generado.

Agradecimientos

Esta investigación ha sido financiada por la Generalitat Valenciana mediante 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 tec-

[3]Conjuntos de sinónimos empleados en WordNet

nologías del lenguaje humano" (TIN2015-65100-R).

Bibliografía

Acharya, S., B. Di Eugenio, A. D Boyd, K. Dunn Lopez, R. Cameron, and G. M Keenan. 2016. Generating summaries of hospitalizations: A new metric to assess the complexity of medical terms and their definitions. In *Proceedings of the 9th International Natural Language Generation conference*, pages 26–30. Association for Computational Linguistics.

Bilmes, J. A. and K. Kirchhoff. 2003. Factored language models and generalized parallel backoff. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology: Companion Volume of the Proceedings of HLT-NAACL 2003–short Papers - Volume 2*, pages 4–6. Association for Computational Linguistics.

Bohnet, B., S. Mille, and L. Wanner. 2011. Statistical language generation from semantic structures. In *Proceedings of the International Conference on Dependency Linguistics*.

Cole, R., J. Mariani, H. Uszkoreit, G. Batista Varile, A. Zaenen, A. Zampolli, and V. Zue. 1997. *Survey of the State of the Art in Human Language Technology*. Cambridge University Press and Giardini.

Ferres, L., A. Parush, S. Roberts, and G. Lindgaard. 2006. Helping people with visual impairments gain access to graphical information through natural language: The igraph system. In *Proceedings of the 10th International Conference on Computers Helping People with Special Needs*, pages 1122–1130. Springer.

Frank, E., M. A. Hall, and I. H. Witten. 2016. *The WEKA Workbench. Online Appendix for "Data Mining: Practical Machine Learning Tools and Techniques"*. Morgan Kaufmann, 4 edition.

Gatt, A., F. Portet, E. Reiter, J. Hunter, S. Mahamood, W. Moncur, and S. Sripada. 2009. From data to text in the neonatal intensive care unit: Using nlg technology for decision support and information management. *AI Commun.*, 22(3):153–186.

Goldberg, E., N. Driedger, and R. I. Kittredge. 1994. Using natural-language processing to produce weather forecasts. *IEEE Expert*, 9(2):45–53.

Jacko, J. A. 2012. *Human-Computer Interaction Handbook: Fundamentals, Evolving Technologies, and Emerging Applications, Third Edition*. CRC Press, Inc., 3rd edition.

Lemon, O., S. Janarthanam, and V. Rieser. 2012. Statistical approaches to adaptive natural language generation. In *Data-Driven Methods for Adaptive Spoken Dialogue Systems*. Springer New York, pages 103–130.

Mann, W. C. and 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.

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

Reiter, E., S. Sripada, J. Hunter, J. Yu, and I. Davy. 2005. Choosing words in computer-generated weather forecasts. *Artificial Intelligence*, 167(1):137–169.

Reiter, E., R. Turner, N. Alm, R. Black, M. Dempster, and A. Waller. 2009. Using NLG to help language-impaired users tell stories and participate in social dialogues. In *Proceedings of the 12th European Workshop on Natural Language Generation*, pages 1–8. Association for Computational Linguistics.

Wan, S., M. Dras, R. Dale, and C. Paris. 2009. Improving grammaticality in statistical sentence generation: Introducing a dependency spanning tree algorithm with an argument satisfaction model. In *Proceedings of the 12th Conference of the European Chapter oof the ACL*, pages 852–860. Association for Computational Linguistics.

Williams, S. and E. Reiter. 2008. Generating basic skills reports for low-skilled readers. *Natural Language Engineering*, 14(04):495–525.

Žolkovskij, A. and I. A. Mel'čuk. 1965. O vozmožnom metode i instrumentax semantičeskogo sinteza. *Naučno-techničeskaja informacija*, 5:23–28.