Modelos secuenciales para enriquecer la tarea de planificación y generación de discurso

Sequence models to boost document planning and discourse generation

Marta Vicente
Universidad de Alicante
mvicente@dlsi.ua.es

Resumen: Como parte de un sistema de generación de lenguaje, la macroplanificación es la fase responsable de proporcionar la estructura adecuada para que el resultado de un proceso de generación produzca un texto coherente. Presentamos una propuesta que incorpora modelos neuronales en un proceso susceptible de captar y producir esa estructura, empleando arquitecturas que han demostrado su utilidad en otros ámbitos del procesamiento del lenguaje.
Palabras clave: Lenguaje natural, macroplanificación, modelos secuenciales

Abstract: In Natural Language Generation, Macroplanning is the stage that implements the plan to produce a meaningful text. Our approach to Macroplanning is based on the adaptation of different neural architectures, that have been found useful in multiple language processing tasks, to examine how they contribute in detecting and providing structure to create discourse.
Keywords: Language generation, macroplanning, sequence models

1 Motivación

En el ámbito de las tecnologías del lenguaje, el área que compete a la generación de lenguaje natural (GLN) incluye la investigación, diseño y realización de artefactos - teorías, técnicas, metodologías, sistemas- cuyo objetivo último es producir texto.

Aunque la disciplina es muy amplia y abarca múltiples niveles de desarrollo, la presente investigación ha limitado su alcance a una de las etapas que constituyen el proceso de generación, la denominada planificación del discurso o macroplanificación.

Dependiendo de la tarea concreta y el objetivo comunicativo que se persiga, el lenguaje natural generado puede adoptar diversas formas. De este modo, la tarea de generar texto carácter a carácter va a plantear un escenario totalmente distinto al requerido por un sistema cuyo cometido sea, por ejemplo, retransmitir un partido de fútbol. En este trabajo, vamos a centrarnos en el tipo de generación que se lleva a cabo cuando el lenguaje toma forma de discurso. Ello se refleja en consideraciones estructurales (superar la frontera de la oración y construir secuencias coherentes de texto) así como semánticas y pragmáticas, dado que, en general, un discurso está adscrito a cierto género textual que, en última instancia, se expresa a través de regularidades que conciernen tanto al tipo de audiencia como a la complejidad léxica o la estructura narrativa.

La decisión relativa a la organización del discurso, en el ámbito de la generación de texto, se ha abordado mediante estrategias que van desde la introducción de reglas o esquemas (Dannélls et al., 2012), pudiendo incluir relaciones entre los elementos del discurso (Williams y Reiter, 2008), hasta la incorporación de técnicas de aprendizaje automático que buscan determinar la ordenación óptima que maximice la coherencia transmitida por el texto generado (Lapata, 2006). Las limitaciones inherentes a estos planteamientos están relacionadas con la dependencia manifiesta respecto al dominio o al género, que reduce la generalización de las aproximaciones y su posible aplicación en condiciones diferentes a aquellas para las que fueron creados.

Dentro de este marco, la investigación aquí reseñada se centra en examinar, aplicar y analizar diferentes técnicas y aproximaciones con el objetivo de detectar las más propicias para construir procesos de macroplanificación que se adapten a circunstancias y

Lloret, E.; Saquete, E.; Martínez-Barco, P.; Moreno, I. (eds.) *Proceedings of the Doctoral Symposium of the XXXIV International Conference of the Spanish Society for Natural Language Processing (SEPLN 2018)*, p. 37–42 Sevilla, Spain, September 19th 2018.

necesidades diversas. Una vez construido el plan del discurso, un módulo de realización que interprete los mensajes allí contenidos generará la salida adecuadamente agregada y flexionada.

2 Descripción de la investigación

La salida de la macroplanificación, por tanto, es el plan de discurso o documento y el objetivo de esta investigación en el momento actual es encontrar el método adecuado para aprenderlo automáticamente, asumiendo que ese plan está asociado a un determinado contexto, que puede estar definido por diferentes elementos como el género, dominio u objetivo comunicativo.

2.1 El plan del documento

En nuestro planteamiento partimos de que el proceso de GLN requiere de una estructura intermedia que dirija y proporcione información relevante a un módulo de realización. Es a esta estructura a la que nos referimos como plan del documento.

El plan de documento estará compuesto por una serie de elementos de información (EI). Un EI es un mensaje preverbal, no una cláusula u oración, y debe poder ser transformado en texto con sentido completo. Esta traducción última la llevaría a cabo el módulo de realización.

Una relación de orden puede existir entre los EI que componen un plan de documento. Su existencia dependerá del tipo de mensaje que se necesite producir. Por ejemplo, es posible que en la reseña de un libro una parte del texto se refiera al argumento del libro (A) empleando varias oraciones (A1,A2,A3) y otra parte a la biografía del autor (B) con sus propias sentencias (B1,B2,B3). Puede que el orden entre A y B no sea relevante para el sentido del texto, mientras que es posible que el orden de los elementos del argumento (A1,A2,A3) sí lo sea.

Si podemos asociar un discurso a un género(s), objetivo comunicativo(s) o tema(s), podemos establecer una relación similar entre estos elementos y el plan de documento así como, en diferente medida, entre estos y un EI. Considerando estas dimensiones como condicionantes del discurso, la adscripción de secuencias de EIs a condicionantes similares influirá en la coherencia del conjunto.

¿De qué modo son relevantes esos elementos en la generación de discurso y en la creación del plan de documento? El género, por ejemplo, puede determinar qué tipo de bloques funcionales deben formar parte del discurso ([planteamiento, nudo, desenlace],[resumen, motivación, metodología, discusión]). Pero también el objetivo comunicativo (entretener, informar) o el tipo de audiencia (academia, niños). En cuanto al tema, determinaría, por ejemplo, si en un cuento se deben introducir animales como personajes o piratas, o aspectos más específicos del espacio narrativo.

3 Metodología propuesta

La metodología propuesta se basa en el análisis y uso de técnicas de aprendizaje para automatizar la extracción y creación de planes de documento, relacionados con un tipo de corpus, que nos permita cierta generalización de los mismos. Se han definido tres objetivos generales:

- Obtener una codificación de cada documento que permita diferenciar y representar las características de sus partes,
- aprender la organización y constitución de tales partes,
- inferir los componentes (EI) del plan de documento.

A partir del análisis de los modelos de lenguaje y considerando la necesidad de superar la limitación que representa un planteamiento basado en bolsa de palabras, nuestra primera aproximación emplea modelos de lenguaje posicionales (Vicente y Lloret, 2017). Basados en métodos de estadística no paramétrica (*Kernel Density Estimation*), son sensibles tanto a la importancia como a la distribución de los elementos en el texto.

En paralelo a este trabajo, se ha comenzado a desarrollar una segunda línea que incorpora las redes neuronales tanto en la tarea de representación de la información como en lo relativo a la secuenciación de las partes del discurso. En lo que sigue, se definen los agentes de tal enfoque, incidiendo en las potencialidades y los retos que conlleva.

3.1 Representación del texto

La primera parte de nuestra aproximación consiste en determinar de qué manera se van a representar las partes del texto y su naturaleza. El propósito es codificar diferentes características del texto, y proporcionar la ex-

presión de su significado en unidades procesables.

Un texto puede ser dividido en varias secciones. Nos referiremos a tales secciones como *ventanas*. La longitud de las mismas, el número de elementos que englobe, dependerá de la técnica empleada. Algunos planteamientos requieren de un numero fijo de elementos en la entrada, otros aceptan un número variable. De ese modo, podemos definir una configuración en la que una ventana contenga siempre 5 palabras frente a otra en la que la ventana coincida con los elementos de cada oración, longitud variable, por tanto.

Cada una de estas ventanas de palabras puede expresarse numéricamente, lo que nos permite emplear tales representaciones en procesamientos posteriores. Los modos de representación que consideraremos en nuestra investigación son:

a) **Codificación *one-hot***. A cada elemento de la ventana, se le asocia un vector del tamaño del vocabulario y en el índice correspondiente, se incluye un valor. Éste puede indicar simplemente la presencia/ausencia del elemento, o puede ser un coeficiente calculado a partir de la frecuencia, el tf-idf, etc. Se determina una técnica para combinarlos.

b) ***Vectores semánticos (word embeddings)***. Cada elemento de la ventana se asocia esta vez a un vector semántico. Los vectores semánticos pueden estar entrenados previamente y adaptados en caso de que el volumen de datos con los que se trabaja sea insuficiente. Por otro lado, estos vectores pueden representar diferentes unidades del texto: palabras (*Word2vec* (Mikolov et al., 2013),*Glove* (Pennington, Socher, y Manning, 2014)), sentido (*Sense2vec* (Trask, Michalak, y Liu, 2015), párrafos (Le y Mikolov, 2014), etc. Se determina una técnica para combinarlos.

c) **Técnicas de *Topic modeling***, como *Latent Direchlt Allocation*(LDA) (Blei, Ng, y Jordan, 2003). Podemos asociar un vector de topics, o relaciones semánticas latentes, a cada ventana del documento empleando LDA, asumiendo de ese modo que tal vector es una representación densa de la misma.

d) **Arquitectura *encoder-decoder***. Esta alternativa, quizá la más interesante en términos de aplicación de redes neuronales, en su modalidad *autoencoder*, implica tomar cualquiera de las anteriores representaciones como entrada y salida de la arquitectura para proporcionar una nueva codificación de la misma, generalmente de menor dimensión. Esa peculiaridad por la que la entrada y la salida coinciden es la razón por la que estos métodos son considerados *semi-supervisados*. El diseño del *autoencoder*, la selección del tipo de red neuronal que se emplee en cada una de sus partes, está condicionado por el hecho de que los elementos de la ventana que debe representar forman una secuencia en la que el orden y las dependencias entre los mismos es relevante. Porque son capaces de modelar tales dependencias, en PLN se suele trabajar con redes recurrentes en alguna de sus modalidades: LSTM, Bi-LSTM, GRU,... Han sido empleados con éxito en lenguaje en tareas como detección de paráfrasis (Socher et al., 2011) o traducción automática (Cho et al., 2014).

Cualquiera sea el tipo de representación seleccionada, cada documento del corpus será definido como una secuencia de las mismas, entre las que se asumirá una relación de orden.

3.2 Modelos de lenguaje para estructurar el discurso

Una de las tendencias en PLN en relación al uso de redes neuronales es la explotación de diferentes tipos de modelos secuenciales con el fin de construir modelos de lenguaje (ML) (Mikolov et al., 2010). Un ML no solo asigna una probabilidad a un conjunto de palabras, sino que permite generar una secuencia de las mismas aplicando, por ejemplo, estrategias de búsqueda sobre un espacio de elementos posibles.

En el caso de la GLN, los ML se han empleado para generar texto carácter a carácter, por ejemplo, pero también para aprovechar las propiedades inherentes a los *word embeddings* como transmisores de significado, de modo que, considerando un vocabulario más extenso que el asociado a un corpus sobre el que se entrena el ML, se puedan generar secuencias en las que aparecen palabras no contenidas originalmente en el corpus de entre-

namiento.

Sin embargo, en cada uno de esos casos, la generación está lejos de crear discurso considerando elementos estructurales. Esto es, la decisión de generar el siguiente elemento, sea éste un carácter o una palabra, está condicionada por la historia inmediatamente anterior al nuevo elemento. Nuestro planteamiento busca trascender esa limitación y para ello aplica un cambio de enfoque.

La estrategia a seguir en esta etapa se basa, por tanto, en los modelos secuenciales pero, en lugar de modelar secuencias de palabras o caracteres, queremos modelar las secuencias de representaciones que definen un documento, tal y como se introdujeron en el apartado 3.1. Nos referiremos a este modelo como *modelo de representaciones*, para diferenciarlo de un modelo de lenguaje *convencional*.

3.3 Generar estructura

Una vez definida la metodología para a) representar un documento junto a su estructura en forma de conjunto de representaciones (Sección 3.1) y b) aprender un modelo sobre secuencias de representaciones (Sección 3.2), la generación de un plan de documento se puede plantear desde dos puntos de vista.

1. **Variaciones del texto original.** Por un lado, partiendo de un texto y las representaciones correspondientes, definiremos métodos para construir planes de documento que conduzcan a variantes del texto original tomándolo como base. Seleccionando un subconjunto de los elementos, conseguiríamos un tipo de resumen y, en la dirección opuesta, el plan de documento podría ser aumentado o enriquecido con otras representaciones, propiciando la realización de una versión extendida del texto original.

2. **Generación libre.** Por otro lado, podríamos emplear el modelo de representaciones entrenado sobre el corpus para generar una secuencia completamente nueva de elementos EI, consiguiendo de ese modo un plan de documento que no coincidiera con ninguno existente en el corpus.

4 Trabajos relacionados

Además de los trabajos mencionados a lo largo del artículo, existe una serie de documentos de referencia que revisan los hitos más importantes de la disciplina de GLN, tanto de sus fundamentos (Reiter y Dale, 2000) como de su estado actual incidiendo en el impacto de las aproximaciones neuronales (Gatt y Krahmer, 2018).

En relación a la aplicación de modelos secuenciales y de arquitecturas *autoencoder* en GLN, algunos trabajo recientes son (Ferreira et al., 2017), usando modelos secuenciales para generar texto desde AMRs (*abstractc meaning representation*) o (Dušek y Jurcıcek, 2016), que adapta la técnica en el ámbito de diálogo.

5 Líneas y cuestiones abiertas

Este trabajo ofrece una propuesta de investigación que se centra en examinar cómo la estructura del discurso toma parte del proceso de generación considerando arquitecturas basadas en, aunque no limitadas por, la incorporación de redes neuronales. Múltiples configuraciones son definidas y cada una propiciará una serie de experimentos con el fin de analizar y determinar cómo captar estructura y cómo producirla cuando el objetivo de la generación es crear discurso.

Las variaciones posibles y la determinación de cada etapa suscita multitud de cuestiones para el debate.
Desde un punto de vista pragmático que considere el discurso como expresión lingüística de un contexto más amplio, ¿cómo codificar, incluir, procesar información a ese nivel más allá de la concerniente a la adscripción a un género y lo que ello comporta?¿En qué manera la comprensión y creación de discurso, en el ámbito de la generación automática, puede verse afectada por las circunstancias en que se produce?

Por otro lado, en relación con la composición y naturaleza de cada representación de información: ¿Es mejor una ventana de palabras, de lemas,... tal vez una composición de diferentes elementos semánticos? y, en ese sentido ¿deberíamos extraer tales características manualmente o debería ser la red neuronal la que las aprendiera?

En relación con el diseño de cada una de las arquitecturas, ¿cuál es más adecuada para cada tarea? ¿Cuánta profundidad, qué número de unidades por capa? O también, ¿es posi-

ble y adecuado combinar el aprendizaje de la representación que determinemos o el modelado de la misma con, por ejemplo, otro tipo de elementos como la polaridad o la emoción asociada a la ventana considerada?

En cuanto a la inclusión de otras técnicas, ¿qué papel podrían jugar aproximaciones como los modelos ocultos de Markov o las técnicas de *topic modelling*? ¿Cómo se integrarían las características latentes del texto procedentes de tales aproximaciones?

Cuestiones éstas que se irán resolviendo en el transcurso de la investigación, desde el estudio, la experimentación y la evaluación.

Agradecimientos

Este proyecto ha sido financiado parcialmente por la Generalitat Valenciana a través del contrato ACIF/2016/501 y la ayuda BEFPI/2018/070, así como el proyecto PROMETEOII/2014/001. También ha participado en su financiación el Gobierno de España a través del proyecto RESCATA (TIN2015-65100-R).

Bibliografía

Blei, D. M., A. Y. Ng, y M. I. Jordan. 2003. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3(Jan):993–1022.

Cho, K., B. van Merrienboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, y Y. Bengio. 2014. Learning phrase representations using rnn encoder–decoder for statistical machine translation. En *Proceedings of the Conference on Empirical Methods on Natural Language Processing*, páginas 1724–1734.

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.

Dušek, O. y F. Jurcıcek. 2016. Sequence-to-sequence generation for spoken dialogue via deep syntax trees and strings. En *Proceedings of the Association for Computational Linguistics*, página 45.

Ferreira, T. C., I. Calixto, S. Wubben, y E. Krahmer. 2017. Linguistic realisation as machine translation: Comparing different mt models for amr-to-text generation. En *Proceedings of the International Conference on Natural Language Generation*, páginas 1–10.

Gatt, A. y E. Krahmer. 2018. Survey of the state of the art in natural language generation: Core tasks, applications and evaluation. *Journal of Artificial Intelligence Research*, 61:65–170.

Lapata, M. 2006. Automatic evaluation of information ordering: Kendall's tau. *Computational Linguistics*, 32:471–484.

Le, Q. y T. Mikolov. 2014. Distributed representations of sentences and documents. En *International Conference on Machine Learning*, páginas 1188–1196.

Mikolov, T., K. Chen, G. Corrado, y J. Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Mikolov, T., M. Karafiát, L. Burget, J. Černockỳ, y S. Khudanpur. 2010. Recurrent neural network based language model. En *Proceedings of the Conference of the International Speech Communication Association*.

Pennington, J., R. Socher, y C. D. Manning. 2014. Glove: Global vectors for word representation. En *Empirical Methods in Natural Language Processing (EMNLP)*, páginas 1532–1543.

Reiter, E. y R. Dale. 2000. *Building natural language generation systems*. Cambridge university press.

Socher, R., E. H. Huang, J. Pennin, C. D. Manning, y A. Y. Ng. 2011. Dynamic pooling and unfolding recursive autoencoders for paraphrase detection. En *Advances in neural information processing systems*, páginas 801–809.

Trask, A., P. Michalak, y J. Liu. 2015. sense2vec - A fast and accurate method for word sense disambiguation in neural word embeddings. *CoRR*, abs/1511.06388.

Vicente, M. y E. Lloret. 2017. Analysing positional language models for natural language generation. En *Proceedings of the Language & Technology Conference 2017*, páginas 357–361.

Williams, S. y E. Reiter. 2008. Generating basic skills reports for low-skilled readers.

Natural Language Engineering, 14(4):495–525.