Introducción a la simulación como metodología de investigación en lingüística

16 de octubre de 2019

1. Introducción

A la hora de obtener datos para nuestras investigaciones, lo más frecuente y accesible suele ser diseñar un experimento conductual, conseguir sujetos y ponerlos a resolver distintas tareas. En estos casos (y en muchos otros), estamos experimentando directamente sobre el *sistema* de interés, definido como un conjunto de *entidades* que actúan e interactúan entre ellas para alcanzar un objetivo. Esto, no obstante, no es siempre posible o eficiente. De ser así, podemos optar por experimentar sobre un *modelo*, es decir, un conjunto de supuestos acerca de cómo funciona ese sistema y que usamos para obtener conocimiento de cómo se comporta ese sistema.

Los *Mars analog habitats* son estaciones de investigación que buscan reproducir las condiciones físicas y psicológicas que los colonizadores experimentarían en un eventual viaje a Marte. De este modo, los investigadores pueden obtener datos útiles para diseñar estas campañas sin la necesidad de experimentar sobre el sistema real, esto es, Marte. A este tipo de modelos se los conoce como *modelos físicos*. Otros ejemplos de modelos físicos pueden involucrar el uso de maquetas, máquinas y otros elementos, que buscan construir dispositivos físicos que recreen las características relevantes del sistema.

En lingüística, pocas veces podemos contruir modelos físicos, especialmente cuando estamos investigando los procesos cognitivos subyacentes a la adquisición o el uso del lenguaje. En estos casos, podemos recurrir a *modelos matemáticos*, es decir, a representaciones del sistema en términos de relaciones lógicas y cuantitativas que pueden ser manipuladas para observar cómo reacciona el modelo. Un ejemplo de estos modelos es la ley de Zipf:

$$f_n \sim \frac{1}{n^a},$$

donde f_n es la frecuencia de un ítem léxico, n es su posición en orden de frecuencia y a es un parámetro mayor o igual a 1. Por ejemplo, si fijamos $a = 1{,}5$, podemos predecir que $f_1 = 1$, $f_2 \approx 0{,}35$ y $f_{10} \approx 0{,}03$. Esto significa que por cada 100 apariciones de la palabra 1, la palabra 2 aparece unas 35 veces y la palabra 10 unas 3 veces. Así, este modelo nos permite obtener resultados y predicciones en forma relativamente sencilla haciendo uso de lápiz, papel y una calculadora. A este tipo de soluciones se las llama *analíticas*.

Pero algunos modelos matemáticos pueden ser más complejos. Por ejemplo, podemos representar el proceso de adquisición del lenguaje como

$$\mathcal{L} : (S_0, E) \to S_T,$$

donde $\mathcal{L}$ es un algoritmo o una función que mapea un estadío inicial de la facultad del lenguaje S_0 con un estadío terminal S_T, sobre la base de la evidencia E presente en un determinado contexto. A partir de esto, se han propuesto distintas definiciones de este algoritmo. Los enfoques empiricistas suelen dar mayor preponderancia a $\mathcal{L}$, mientras que

los enfoques racionalistas la otorgan a S_0. Encontrar una solución analítica a los modelos propuestos suele ser complejo (aunque para algunos de ellos podríamos, en principio, obtenerla con ayuda de una computadora). No obstante, una alternativa consiste en correr una *simulación*, es decir, en alimentar el modelo con inputs para observar cómo se ven afectadas las medidas de rendimiento del output.

2. Modelos de simulación discretos

El sistema que una simulación define puede presentar diversos *estados*. Estos estados se definen como la colección de todas las variables necesarias para describir el sistema en un tiempo particular. En las simulaciones discretas, el sistema solo puede cambiar de estado en ciertos momentos, esto es, las variables de estado no se actulizan constantemente. A estos momentos los llamamos *eventos*: ocurrencias instantáneas que *pueden* cambiar el estado del sistema. Una simulación de estados discretos va a estar conformada por los siguientes elementos:

- un *estado del sistema*: la colección de variables de estado necesarias para describir el sistema en un momento particular
- un *reloj de simulación*: una variable que realiza el seguimiento del tiempo simulado
- una *lista de eventos*: una lista que contiene los momentos en los que los distintos tipos de eventos ocurrirán
- *contadores estadísticos*: variables usadas para almacenar información estadística sobre el rendimiento del sistema
- una *rutina de inicialización*: se ocupa de inicializar las variables del modelo en el tiempo 0
- una *rutina de sincronización*: determina el próximo evento y avanza el reloj de simulación a ese punto
- *rutinas de evento*: se ocupan de definir cómo los distintos eventos alterarán el estado del sistema
- *rutinas de librería*: se ocupan de generar observaciones aleatorias a partir de distribuciones de probabilidad definidas como parte del modelo
- un *generador del reporte*: produce el reporte cuando la simulación termina
- un *programa principal*: se ocupa de coordinar los distintos elementos de la simulación

3. Un ejemplito psicolingüístico

Para ver cómo podemos usar todo esto para obtener conocimiento acerca del funcionamiento del lenguaje vamos a correr una simulación del proceso de adquisición del lenguaje. Esta simulación fue diseñada *solo* con fines ilustrativos a partir de simplificaciones sobre los modelos propuestos en la bibliografía, y no se propone como un experimento válido, sino como una mera muestra del diseño y la lógica de una simulación usada como metodología de investigación.

3.1. El sistema y el modelo

El sistema que queremos estudiar es el conocimiento que un hablante tiene de su lengua. En particular, queremos saber cómo se adquiere ese conocimiento. Para esto vamos a necesitar un modelo matemático de este proceso. Habíamos dicho que, en términos generales, podemos representar la adquisición de una lengua como una función o algoritmo $\mathcal{L}$ que mapea un estado inicial s_0 con un estado terminal s_T a partir de la exposición a evidencia lingüística E:

$$\mathcal{L} : (S_0, E) \to S_T.$$

Específicamente, vamos a considerar el modelo propuesto por Yang (2002), quien propuso que la adquisición de una lengua es el producto de la "competencia" (en el sentido de competir) entre todas las posibles gramáticas disponibles al niño por la Gramática Universal:

$$G_1, G_2, ..., G_n.$$

Cada una de estas gramáticas estaría asociada a un peso $(p_1, p_2, ..., p_n)$. Ante un determinado estímulo lingüístico, el niño samplearía una gramática en forma probabílistica en función del peso que tiene asociada. Inmediatamente los pesos de las gramáticas se actualizan de acuerdo al éxito o fracaso de dicha gramática para parsear el estímulo en cuestión. Formalmente, podemos modelar esta actualización a partir del esquema lineal de recompensa-castigo:

$$G_i \to s \Rightarrow \begin{cases} p_i' = p_i + \gamma(1 - p_i) \\ p_j' = (1 - \gamma)p_j, & \text{con } j \neq i \end{cases}$$

$$G_i \nrightarrow s \Rightarrow \begin{cases} p_i' = (1 - \gamma)p_i \\ p_j' = \dfrac{\gamma}{N - 1} + (1 - \gamma)p_j, & \text{con } j \neq i \end{cases}$$

donde $G_i \to s$ expresa que G_i genera/interpreta el estímulo lingüístico s, $G_i \nrightarrow s$ expresa que G_i no lo genera/interpreta, y γ es un parámetro.

3.2. El input

El input para esta simulación va a ser de dos tipos: un conjunto de gramáticas posibles; y una lista de estímulos lingüísticos. Vamos a representar el conjunto de gramáticas como:

$$G = \{G_{\text{neerlandés}}, G_{\text{inglés}}, G_{\text{irlandés}}, G_{\text{hixkaryana}}\},$$

donde $G_{\text{neerlandés}}$ es una gramática que acepta los órdenes de palabras del neerlandés, $G_{\text{inglés}}$ es una gramática que acpeta los órdenes de palabras del inglés, y así sucesivamente. Específicamente,

- Neerlandés: SVO, XVSO, OVS
- Inglés: SVO, XVSO
- Irlandés: VSO, XVSO
- Hixkaryana: OVS, XOVS

Los estímulos lingüísticos, los vamos a generar probabilísticamente en función de la proporción en que están presentes en el contexto al que es expuesto un niño que adquiere neerlandés, esto es, 64.7 % SVO, 34 % XVSO y1,3 % OVS, aproximadamente.

3.3. El output

El output de nuestra simulación va a ser un vector:

$$\vec{P} = (p_{\text{neerlandés}}, p_{\text{inglés}}, p_{\text{irlandés}}, p_{\text{hixkaryana}}),$$

donde cada componente representa un peso asociado a uno de los tipos de gramática y la suma de las componentes es siempre igual a 1.

4. Ventajas y desventajas

4.1. Ventajas

- Las simulaciones son el único método de investigación disponible para la mayoría de sistemas complejos, que no pueden ser descritos precisamente por un modelo matemático que pueda ser evaluado analíticamente.
- Las simulaciones nos permiten estimar el rendimiento de sistemas existentes bajo condiciones nuevas.
- Las simulaciones nos permiten tener un mejor control de las condiciones experimentales del que podemos tener cuando experimentamos directamente con el sistema.
- Las simulaciones nos permiten comprimir o expandir el tiempo del estudio.

4.2. Desventajas

- Cada ejecución de un modelo de simulación estocástico solo nos permite obtener estimaciones de las verdaderas características del modelo bajo determinados input, mientras que los modelos analíticos pueden frecuentemente producir las características verdaderas exactas del modelo para una variedad de conjuntos de parámetros de entrada.
- Los modelos de simulación pueden ser caros y de desarrollo lento.
- Los resultados cuantitativos o impresionistas de las simulaciones pueden dar lugar a una sobreconfianza injustificada en el modelo. Si el modelo no es una representación válida del sistema bajo estudio, los resultados de la simulación proveerán información poco útil sobre el sistema en cuestión.