Ciencia de Datos y Lenguaje Natural

Teoría de la Información - 1

Grupo PLN - INCO

Universidad de la República

Qué es la teoría de la información

Teoría matemática que se ocupa de:

- Comunicación
- Compresión de datos
- Complejidad, correlacción, incertidumbre
- Una teoría matemática que proporciona métricas para cuantificar grados de divergencia y correlación entre distribuciones.

Teoría de la información y lenguaje

- Basada en probabilidades, no no favorece las categorías rígidas.
- Y hay variados fenómenos del lenguaje que admiten respuestas graduables.

Teoría de la información y lenguaje

- El verbo *comer* requiere un objeto que sea comestible, pero se dice *se come las eses cuando habla*
- una hemorragia es una pérdida de sangre, pero se dice *el país no puede soportar esa hemorragia financiera*
- La metáfora se da permanentemente en el lenguaje corriente, es un deslizamiento del significado.

Teoría de la información y lenguaje

The Category Squish: Endstation Hauptwort*

John Robert Ross

M.I.T.

In this paper, I will examine a number of phenomena which suggest that the traditional distinction between verbs, adjectives, and nouns--a distinction which is commonly thought of as discrete--should be modified. I will postulate, instead of a fixed, discrete inventory of syntactic categories, a quasi-continuum, which contains at least the categories shown in (1), ordered as shown there.

(1) Verb > Present participle > Perfect participle > Passive participle > Adjective >

Restricciones de selección

Las restricciones o preferencias de selección expresan como los verbos restringen el tipo semántico de los argumentos:

- *Una soprano cantó el Aleluya* vs.
- *Una bicicleta cantó un martillo*

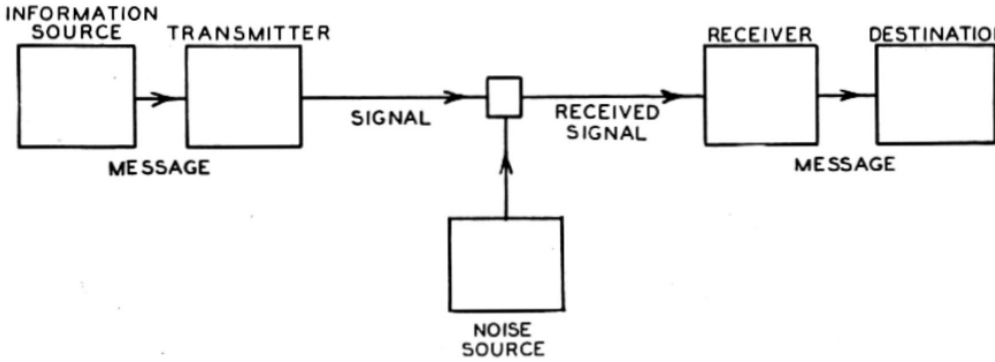
`https://web.stanford.edu/class/psych227/resnik.pdf`

Restricciones de selección

- Resnik (1993) define una asociación predicado-argumentos de modo probabilístico, como la cantidad de información que un predicado expresa acerca de la clase semántica de los argumentos.

Verb	Direct Object Semantic Class	Assoc	Direct Object Semantic Class	Assoc
read	WRITING	6.80	ACTIVITY	-.20
write	WRITING	7.26	COMMERCE	0
see	ENTITY	5.79	METHOD	-0.01

¿Teoría de la información o de la comunicación?



Shannon (1948). A mathematical theory of communication. Bell System Technical Journal 27(3): 379—423.

¿Teoría de la información o de la comunicación?

The fundamental problem of communication is that of reproducing at one point either exactly or approximately a message selected at another point. Frequently the messages have meaning; that is they refer to or are correlated according to some system with certain physical or conceptual entities. These semantic aspects of communication are irrelevant to the engineering problem.

Shannon (1948). A mathematical theory of communication. Bell System Technical Journal 27(3): 379—423.

¿Teoría de la información o de la comunicación?

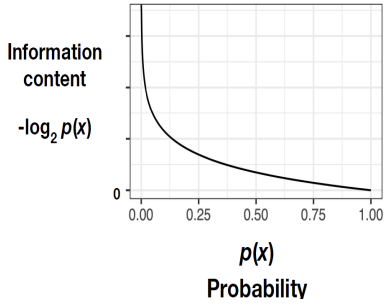
- Si bien no se considera lo esencial del mensaje, el contenido, el enfoque resulta interesante para tareas de lenguaje.
- La noción de contenido de información expresada como información nueva, se adecua bien a diversos problemas.

Contenido de información

- Supongamos un evento x que ocurre con determinada probabilidad p.
- Cuánto menor sea la probabilidad de ocurrencia, mayor será la "sorpresa" que genere, o sea, la información que aporte.
- Se define la sorpresa o contenido de información como

$$ci(x) = -log_2 p(x).$$

Contenido de información



- El valor mínimo, o sea 0, lo toma cuando el evento es seguro, o sea , la probabilidad es 1.
- Es una función decreciente en el intervalo (0, 1]

Contenido de información

Shannon definió la función de modo que cumpla los siguientes axiomas:

1. La ocurrencia de un evento con probabilidad 1 no aporta ninguna información; su contenido de información es 0.
2. Cuanto menos probable es un acontecimiento, más información aporta.
3. La cantidad total de información de dos eventos independientes es la suma de las cantidades de información de los eventos individuales. *Esto se conoce como una propiedad de monotonía en lógica. No es necesariamente cierto: información nueva puede invalidar algo previo.*

Contenido de información

1. Es posible demostrar que la función $ci(x) = -log_2 p(x)$ cumple con las propiedades anteriores.
2. La aditividad buscada se asegura con el logaritmo. Para dos eventos independientes la probabilidad de que ocurran ambos es el producto de las probabilidades y el contenido de información es la suma de los contenidos.
3. Es posible demostrar también que si se cumplen los requisitos previos, la función de contenido de información tiene la forma $ci(x) = -log_b p(x)$
   O sea, la función que hemos definido previamente, para cualquier base b.

Entropía

Se define la entropía como:

$$H(X) = -\sum_{i=1}^{n} p(x_i) \log_2 p(x_i).$$

- Es la esperanza de la función ci (contenido de información).
- El contenido de información se asocia a un evento, o sea, un valor posible de una variable aleatoria.
- La entropía se asocia a la variable aleatoria: el conjunto de valores que puede tomar y sus probabilidades.

Ejemplo: contenido de información, entropía

i	a_i	p_i	$h(p_i)$
1	a	.0575	4.1
2	b	.0128	6.3
3	c	.0263	5.2
4	d	.0285	5.1
5	e	.0913	3.5
6	f	.0173	5.9
7	g	.0133	6.2
8	h	.0313	5.0
9	i	.0599	4.1
10	j	.0006	10.7
11	k	.0084	6.9
12	l	.0335	4.9
13	m	.0235	5.4
14	n	.0596	4.1
15	o	.0689	3.9
16	p	.0192	5.7
17	q	.0008	10.3
18	r	.0508	4.3
19	s	.0567	4.1
20	t	.0706	3.8
21	u	.0334	4.9
22	v	.0069	7.2
23	w	.0119	6.4
24	x	.0073	7.1
25	y	.0164	5.9
26	z	.0007	10.4
27	-	.1928	2.4
$\sum_i p_i \log_2 \frac{1}{p_i}$			4.1

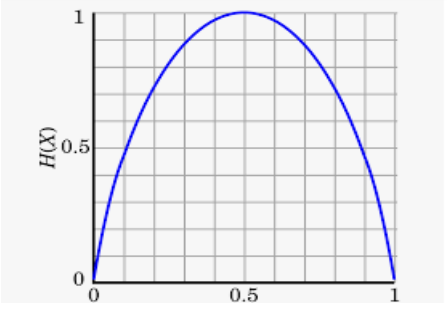
Propiedades de la entropía

- Es siempre positiva.
- Dada una variable aleatoria discreta $X = x_1, ..., x_n$, donde cada uno de los valores x_i tiene probabilidad p_i, $H(X) = -\sum_{i=1}^{n} p(x_i) \log_2 p(x_i)$ es máxima cuando los valores x_i son equiprobables

Propiedades de la entropia

La entropía es siempre positiva.

El valor maximo se obtiene para eventos equiprobables.



Entropía en función de la probabilidad p variable de cara al tirar una moneda

Variantes de la entropía

- Se definen cantidades vinculadas con la entropía, que se usan en la resolucion de diversos problemas.
- En general implican más de una variable, porque se trata de comparar distribuciones.