

Estimación de la confiabilidad del modelo de MR para el análisis de AI en español

Ken Matsuda Oteiza

Confiabilidad Inter Codificadores

Con el propósito de validar el modelo de Movidas Retóricas propuestas en este proyecto se tomó una muestra aleatoria de 10 textos del corpus, a modo de corpus de entrenamiento. Este subcorpus estaba compuesto por 10 artículos de investigación seleccionados aleatoriamente, según los estratos del CaiE, es decir, tiene la misma distribución disciplinar que el CaiE, y por lo tanto, que Scielo.

En el Cuadro 1, se presenta el corpus de entrenamiento:

ID	Nombre	Disciplina	Revista
1	Priming sintáctico en la producción de oraciones en español. Acerca de la representación sintáctica de los verbos	Lingüística	RLA
2	Análisis de la presencia de actinomicosis pélvica en mujeres de una comunidad rural en Chile	Obstetricia/ginecología	Revista chilena de obstetricia y ginecología
3	Implementación de un reconocedor de palabras aisladas dependiente del locutor	Ingeniería	REVISTA FACULTAD DE INGENIERÍA, Cuadernos de economía
4	El dinero como indicador líder	Economía	Estudios pedagógicos (Valdivia)
5	Revalidación de constructor índice de calidad del aprendizaje profesional	Educación	
6	Prodigios que abruma: Dos cuentos de Rubén Darío	Literatura	Acta literaria
7	Los usos políticos de la memoria y la identidad	Antropología	Estudios atacameños

8			Revista chilena de enfermedades respiratorias
	Caso clínico-radiológico adulto	Medicina	
9	Adrenalectomía laparoscópica: lecciones aprendidas en 110 procedimientos consecutivos	Cirugía	Revista chilena de cirugía
10	Efecto del arreglo de racimo y aplicación del citoquinina sintética (CPPU) en la calidad de uva de mesa variedad sultanina tratada con dos fuentes de giberelinas	Agricultura	Agricultura Técnica

Cuadro 1: Corpus de entrenamiento.

Cada uno de estos textos, fue guardado en formato RTF y fue traspasado al software Atlas ti, versión 5.1 . Se realizaron dos sesiones de instrucción en el uso de este software con 6 codificadores que previamente habían sido entrenados en el uso del libro de códigos. Este libro de códigos está compuesto por un listado de 36 MR. Cada código está asociado a una categoría definida conceptualmente.

Finalizada la codificación para calcular la confiabilidad Inter- Codificadores (Inter Rater Reliability), se utilizó un software específico: CAT (Coding Análisis Toolkit) que fue desarrollado bajo licencia GLP (General Public License) en la Universidad de Pittsburg en su programa QDAP (Qualitative Data Análisis Program) en el año 2007. Este es una aplicación específica que permite, entre otras funcionalidades, el cálculo de la confiabilidad entre codificadores, que han marcado o codificado textos con la aplicación Atlas.ti. Es posible elegir entre el alfa de Krippendorff o el Kappa de Fleiss. En este proyecto se ha elegido el Kappa de Fleiss

El Kappa de Fleiss

Es una generalización de la pi de Scott(1955), una medida estadística de fiabilidad o consistencia interna. También relacionado con el estadístico Kappa de Cohen(1968), tanto el de Scott como el de Cohen(1973) solamente permiten calcular la confiabilidad solo con dos jueces o codificadores en nuestro caso.

El Kappa de Fleiss (1971, 1981) en cambio, permite calcular la confiabilidad con dos o más jueces, y para una cantidad fija de elementos. Puede ser interpretada como la cantidad observada de acuerdo entre codificadores que excede lo que sería de esperar si todos los codificadores hiciesen sus codificaciones aleatoriamente.

La fórmula para el cálculo de kappa es:

$$\kappa = \frac{\bar{P} - \bar{P}_0}{1 - \bar{P}_0}$$

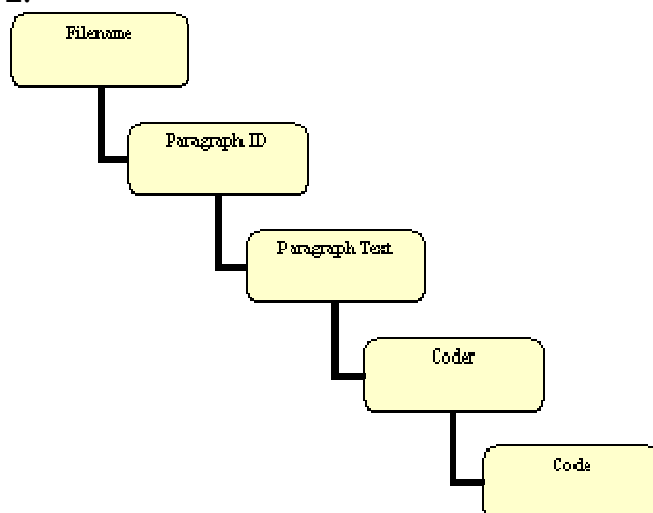
Donde el numerador representa el grado o la probabilidad de acuerdo alcanzado realmente descontando la proporción o probabilidad de acuerdos ocurrida por azar y el denominador es el grado o probabilidad total de acuerdos posibles descontando los que se han producido por azar.

Algoritmo utilizado por el sistema CAT

En el algoritmo utilizado por CAT, un valor kappa se calcula para cada código entre los codificadores que ingresaron a ese código para cualquier unidad. A nivel interno, lo que sucede durante la fase de cómputo en CAT es lo siguiente:

1. En un paso preliminar, se crea una estructura interna de datos, a partir de los datos codificados que contiene la siguiente jerarquía: En primer lugar se considera el nombre del documento primario, en segundo lugar se numeran los párrafos, en tercer lugar el texto que contiene el párrafo, en tercer lugar se identifica al codificador y finalmente el código que asignó el codificador. Simultáneamente se genera, para cada código, un vector que contiene los elementos señalados precedentemente y con estos vectores se crea una lista de los codificadores y la codificación a un párrafo específico.

2.



3. Para cada código presente en el conjunto de datos, el sistema itera sobre la lista de vectores (de nombre de archivo, el apartado de identificación y texto) que se creó anteriormente y genera el conteo de los códigos que coinciden exactamente (SÍ), y de las coincidencias parciales de los códigos, además de los que no coinciden (NO), y la cuenta total para el código. En síntesis, en un "sí" el valor es un código que un codificador ha asignado a una unidad de código, un "no" cuando el valor es un codificador que ha codificado una unidad, pero no teniendo en cuenta que la unidad pertenece a un código particular. Como un paso opcional, el CAT, realiza una comprobación para ver si la codificación de los párrafos del texto se superpone y crea un nuevo conjunto de vectores para estos registros

Finalmente, una vez que todas las estructuras de datos se configuran dentro del CAT, el cálculo de Kappa se hace en varios pasos:

- a. Calcular el valor de $c * (c-1)$ como "NN1" (donde c es el número de programadores) - también calcular la inversa del valor como $1/NN1$

- b. Luego, para cada código, obtener el "sí" y "no" cuenta como derivados del paso 2
- c. Calcular la probabilidad de un recuento de sí como (# sí / número total) y un recuento no como (# sin / número total)
- d. A continuación, para cada tupla (vector) pertenecientes a este código la estadística para Pise calcula como:

$$\pi = \sum_{i=1}^{N_{tuplas}} \frac{(Count_{no})^2 + (Count_{yes})^2 - (Count_{coders})^2}{NN1}$$

- e. Además, si los códigos se superpone se calcula, la estadística de PI en los solapamientos se calculan así:

$$\pi = \sum_{i=1}^{N_{tuplas}} \frac{(Count_{no})^2 + (Count_{yes})^2 + (Count_{overlap})^2 - (Count_{coders})^2}{NN1}$$

- f. Usando estas sumas y una generalización de la ecuación kappa, el kappa se calcula para cada código, como:

$$\kappa = \frac{\pi}{N_{tuplas}}$$

Una vez calculados los Kappas individuales, el Kappa final es el promedio de ello.

Interpretación de Kappa

Landis y Koch (1977) proporcionaron algunos indicadores de carácter empírico para la interpretación de los valores de κ . Sin embargo, estos no son de ninguna manera universalmente aceptados. Esto es, porque los autores mencionados se basaron en su opinión personal. Según Sim, J. and Wright, C. C. (2005) el número de categorías y temas que afectan a la magnitud del valor. El kappa será mayor cuando hay menos categorías.

κ	Interpretation
< 0	Poor agreement
0.0 – 0.20	Slight agreement
0.21 – 0.40	Fair agreement
0.41 – 0.60	Moderate agreement
0.61 – 0.80	Substantial agreement
0.81 – 1.00	Almost perfect agreement

Resultados y conclusión

Los cálculos realizados muestran que el valor de K total de este modelo es de **0,64** y el valor de K para considerando el solapamiento entre los codificadores es de **0,68**, en ambos casos, se trata de un acuerdo sustancial.

Este procedimiento valida empíricamente este modelo.

A continuación, en la Tabla 1, se exponen los resultados del Kappa específicos para cada código por codificador:

Code	Cristina n Pérez	Deniss e	Gabriel a	Miguel Angel Valenzuela	Miguel Fuente	Paulin Toro	Exact Matc h	Partia Matc h	Overla p	Kapp a	Kappa (inc. Overlap)
1	20	29	18	17	24	18	4	19	9	0.64	0.69
10	14	3	5	1	9	4	0	5	2	0.63	0.65
11	4	4	3	4	1	2	0	5	1	0.55	0.60
12	81	4	4	2	6	4	0	5	2	0.66	0.67
13	5	2	6	2	4	2	0	1	0	0.66	0.66
14	0	0	6	2	5	1	0	1	0	0.65	0.65
15	17	102	28	12	33	19	0	16	30	0.65	0.74
16	0	2	0	0	3	4	0	0	0	0.67	0.67
18	0	3	8	2	6	5	0	3	1	0.64	0.65
19	7	0	2	3	3	2	0	2	0	0.63	0.63
2	19	8	8	7	3	7	0	11	3	0.60	0.64
20	48	4	16	3	12	1	0	6	6	0.65	0.68
21	3	0	24	2	8	0	0	0	3	0.67	0.70
22	17	12	51	16	6	3	0	6	5	0.65	0.67
23	54	86	73	45	43	30	0	43	112	0.63	1.02
24	10	1	35	1	0	2	0	0	0	0.67	0.67
25	14	18	0	15	11	9	0	14	9	0.60	0.71
26	7	5	0	6	6	4	0	6	2	0.59	0.70
27	2	0	1	0	0	0	0	0	0	0.67	0.67
28	1	0	4	0	0	0	0	1	0	0.62	0.62
29	5	11	1	2	5	3	0	3	1	0.63	0.65
3	62	76	71	56	62	33	1	67	93	0.61	0.93
30	3	0	1	3	3	1	0	0	1	0.67	0.70
31	14	1	8	4	4	4	0	4	3	0.64	0.67
32	86	85	79	42	35	17	0	65	107	0.61	0.92
33	19	20	7	3	12	0	0	5	5	0.65	0.69
34	1	1	2	1	8	7	0	2	0	0.64	0.65
35	2	0	1	0	0	0	0	0	0	0.67	0.67
36	2	0	1	1	0	0	0	0	0	0.67	0.67
37	10	1	0	0	0	0	0	0	0	0.67	0.67
38	0	18	5	0	0	0	0	0	0	0.67	0.67
39	2	0	33	0	0	0	0	1	0	0.66	0.66
4	7	11	12	10	10	8	0	10	6	0.61	0.66
5	62	96	41	46	19	64	1	36	62	0.64	0.81
6	6	7	6	4	8	2	0	7	2	0.60	0.63
7	64	106	66	29	30	14	0	40	46	0.63	0.74
8	69	120	93	106	44	24	0	77	137	0.62	0.99
9	8	3	6	1	6	4	0	4	2	0.62	0.65
a emplear en el trabajo por medio de	0	0	0	0	1	0	0	0	0	0.67	0.67

Proyecto Fondecyt 11080097: La variación del artículo de investigación a través de las disciplinas: El caso del indexador SciELO Chile

tablas adelanta lo que realizará después	0	0	1	0	0	0	0	0	0	0.67	0.67
Analiza y/o interpreta el cuento	0	0	0	0	0	15	0	0	0	0.67	0.67
Caracteriza al paciente	0	0	0	0	0	1	0	0	0	0.67	0.67
Conclusión	5	6	7	3	3	3	0	4	2	0.65	0.70
Conclusión de AI	2	2	0	3	3	4	0	3	1	0.60	0.64
Condiciona los resultados ante otras variables	0	0	0	0	1	0	0	0	0	0.67	0.67
Da a entender que los resultados obtenidos hay que mirarlos con cuidado	0	0	0	0	2	0	0	0	0	0.67	0.67
Define conceptos médicos	0	0	0	0	0	3	0	0	0	0.67	0.67
Desarrollo de AI	1	0	0	4	4	1	0	0	0	0.67	0.68
Determina la función de un relato	0	0	0	0	1	0	0	0	0	0.67	0.67
Discusión	5	4	4	4	4	3	0	5	2	0.60	0.66
Ejemplifica con un verso	0	0	0	0	10	0	0	0	0	0.67	0.67
Ejemplifica el análisis con un extracto de otra obra	0	0	0	0	0	2	0	0	0	0.67	0.67
Ejemplifica el análisis con un extracto del cuento	0	0	0	0	0	9	0	0	0	0.67	0.67
ejemplifica una teoría previa	0	0	1	0	0	0	0	0	0	0.67	0.67
el autor presenta evidencia textual de lo analizado	21	0	0	0	0	0	0	0	0	0.67	0.67
establece una perspectiva metodológica desde la cual realiza la investigación	0	0	1	0	0	0	0	0	0	0.67	0.67
evalúa los resultados	0	0	2	0	0	0	0	0	0	0.67	0.67
evalúa resultados de la propia investigación	0	0	1	0	0	0	0	0	0	0.67	0.67
Explicita el motivo de la consulta	0	0	0	0	0	1	0	0	0	0.67	0.67
explicita la conclusión de la investigación	0	0	1	0	0	0	0	0	0	0.67	0.67

Proyecto Fondecyt 11080097: La variación del artículo de investigación a través de las disciplinas: El caso del indexador SciELO Chile

Explicita los procedimientos realizados	0	0	0	0	0	1	0	0	0	0.67	0.67
Explicita los resultados del examen físico y/o de laboratorio	0	0	0	0	0	3	0	0	0	0.67	0.67
Fórmulas	0	15	0	0	0	0	0	0	0	0.67	0.67
Indica el tratamiento	0	0	0	0	0	2	0	0	0	0.67	0.67
Indica la metodología para llegar al diagnóstico	0	0	0	0	0	1	0	0	0	0.67	0.67
Indica respuesta al tratamiento	0	0	0	0	0	1	0	0	0	0.67	0.67
Interpreta los resultados de los exámenes	0	0	0	0	0	1	0	0	0	0.67	0.67
Introducción de AI	8	7	10	5	5	5	0	6	2	0.66	0.69
Marco Teórico	1	3	0	3	5	5	0	3	1	0.61	0.64
Metodología	2	3	5	2	1	1	0	3	0	0.62	0.63
Notas	6	5	7	5	5	5	0	5	2	0.67	0.70
Notas de AI	3	2	5	2	2	2	0	2	1	0.67	0.69
Palabras clave en español	2	1	0	1	3	3	0	2	1	0.60	0.68
Palabras clave en español de AI	6	7	5	6	5	5	0	5	2	0.65	0.70
plantea una pregunta	1	1	3	1	3	2	0	4	1	0.48	0.56
Realiza un análisis semiótico de un poema	0	3	0	0	0	0	0	0	0	0.67	0.67
Realiza un análisis semiótico de un poema	0	0	0	0	12	0	0	0	0	0.67	0.67
Referencias bibliográficas	0	0	0	0	1	0	0	0	0	0.67	0.67
Referencias bibliográficas de AI	8	7	4	5	5	5	0	7	3	0.59	0.64
Resultados y Discusión	1	3	5	5	4	4	0	3	4	0.63	0.75
Resume el cuento	6	6	5	4	4	4	0	5	2	0.62	0.68
Resumen en español	2	0	1	2	1	1	0	2	0	0.57	0.57
Resumen en español de AI	0	0	0	0	0	2	0	0	0	0.67	0.67
Separa en etapas el desarrollo de un poema	6	7	5	5	5	5	0	5	2	0.67	0.70
Título de AI	1	2	3	3	3	3	0	3	0	0.59	0.60
utiliza pregunta retórica	0	0	0	0	1	0	0	0	0	0.67	0.67
Título	8	7	0	5	5	5	0	7	2	0.59	0.64
Título de AI	1	2	9	4	3	4	1	3	0	0.65	0.65
utiliza pregunta retórica	0	0	1	0	0	0	0	0	0	0.67	0.67

Totales 841 932 811 520 531 410 7 542 678 0.64 0.68
Tabla 1: Inter rater reliability → K de Fleiss

Bibliografía

- Fleiss, J. L. (1971) "Measuring nominal scale agreement among many raters." *Psychological Bulletin*, Vol. 76, No. 5 pp. 378–382
- Fleiss, J. L. (1981) *Statistical methods for rates and proportions*. 2nd ed. (New York: John Wiley) pp. 38–46
- Fleiss, J. L. and Cohen, J. (1973) "The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability" in *Educational and Psychological Measurement*, Vol. 33 pp. 613–619
- Gwet, K. (2001) *Statistical Tables for Inter-Rater Agreement*. (Gaithersburg: StatAxis Publishing)
- Gwet, K. L. (2008) "Computing inter-rater reliability and its variance in the presence of high agreement", *British Journal of Mathematical and Statistical Psychology*, Vol. 61, pp29–48
- Gwet, K. L. (2010) *Handbook of Inter-Rater Reliability* (2nd Edition). (Gaithersburg : Advanced Analytics, LLC) ISBN 978-0970806222
http://en.wikipedia.org/wiki/Fleiss%27_kappa(consultado 29/9/2010)
- Landis, J. R. and Koch, G. G. (1977) "The measurement of observer agreement for categorical data" in *Biometrics*. Vol. 33, pp. 159–174
- Scott, W. (1955). "Reliability of content analysis: The case of nominal scale coding." *Public Opinion Quarterly*, Vol. 19, No. 3, pp. 321–325.
- Sim, J. and Wright, C. C. (2005) "The Kappa Statistic in Reliability Studies: Use, Interpretation, and Sample Size Requirements" in *Physical Therapy*. Vol. 85, No. 3, pp. 257–268
- Cohen, J. (1968). "Weighed kappa: Nominal scale agreement with provision for scaled disagreement or partial credit". *Psychological Bulletin* 70: 213–220. [doi:10.1037/h0026256](https://doi.org/10.1037/h0026256)(consultado 29/9/2010)