Procesamiento del Lenguaje Natural, núm. 43 (2009), pp. 65-73 recibido 1-05-2009; aceptado 5-06-2009

Bases para evaluar la anotación de corpus de emociones espontáneas*

On the assessment of the annotation of spontaneous emotional corpora

Zoraida Callejas, Ramón López-Cózar
Dpto. Lenguajes y Sistemas Informáticos. Universidad de Granada
C/ Pdta. Daniel Saucedo Aranda, 18071, Granada
{zoraida, rlopezc}@ugr.es

Resumen: En este artículo proponemos un marco en el que basar la evaluación de las anotaciones de corpus de emociones espontáneas, que emplea diversos coeficientes para medir el acuerdo entre anotadores y la complejidad de la anotación. Además, tiene en cuenta información contextual para interpretar los valores obtenidos. Nuestros resultados experimentales usando un corpus de emociones espontáneas muestran que las interpretaciones tradicionales que encontramos en la literatura no son válidas en el caso de emociones no actuadas, pues el hecho de que la categoría neutra sea predominante por naturaleza conlleva que estas interpretaciones sean poco informativas o incluso contradictorias. Nuestras propuestas proveen de mecanismos sobre los que sustentar una interpretación más fiable de la aceptabilidad de estos corpus.
Palabras clave: Anotación de corpus, emociones no actuadas, sistemas de diálogo

Abstract: In this paper we propose the use of several statistical coefficients and information sources to create an appropriate basis for the evaluation of the annotations of non-acted emotional corpora. Experimental results over a corpus of spontaneous emotions show that traditional interpretations that can be found in the literature are not valid for this type of corpora in which the neutral category is inherently predominant. Our proposals provide sufficient information to obtain reliable interpretations of acceptability
Keywords: Corpus annotation, non-acted emotions, dialogue systems

1. Introducción

Debido a sus beneficios y a su gran variedad de aplicaciones, la computación afectiva ha emergido como una línea de investigación puntera en el campo de la interacción persona-ordenador. Usualmente, los trabajos en esta materia están basados en corpus emocionales que se emplean para entrenar los métodos propuestos y obtener resultados experimentales. En la literatura hay tres enfoques principales para la adquisición de corpus emocionales: grabar habla espontánea, grabar emociones inducidas, y usar actores que simulen las emociones.

Como algunos autores han indicado, p.ej. Douglas-Cowie et al. (2003), la relación entre las emociones actuadas y las espontáneas no se conoce de forma exacta. Sin embargo, como ya indicaba Johnstone (1996), incluso el habla actuada profesionalmente pierde realismo, pues existen algunos efectos que no pueden ser controlados de forma consciente. Diferentes estudios han mostrado que no es apropiado emplear datos actuados para reconocer emociones que se dan de forma natural (Vogt y André, 2005)(Wilting, Krahmer, y Swerts, 2006).

Por tanto, el habla emocional espontánea, que refleja la producción completamente natural de la emoción dentro del dominio de aplicación, es el enfoque más realista. Sin embargo, se necesita un esfuerzo considerable para anotar el corpus, puesto que requiere que para cada grabación se interprete qué emoción ha sido expresada. Emplear un buen esquema de anotación es esencial, pues afecta al resto de las etapas del proceso de aprendizaje. Además, la anotación manual de

* Esta investigación ha sido financiada por el Proyecto HADA TIN2007-64718 del Ministerio de Educación y Ciencia.

 © 2009 Sociedad Española para el Procesamiento del Lenguaje Natural

corpus es muy difícil, requiere mucho tiempo y es costosa, por tanto debe ser diseñada y evaluada cuidadosamente

Sin embargo, en los estudios de evaluación de anotaciones no se suele tener en cuenta las características propias de los corpus de emociones espontáneas, lo que hace que las interpretaciones obtenidas no sean fiables. Esto se debe a que es necesario un esfuerzo muy importante para anotar estos corpus, lo cual está motivado por dos razones. En primer lugar porque están inherentemente desequilibrados al ser los estados neutros mucho más frecuentes que los comportamientos claramente emocionales. En segundo lugar, porque las emociones son más sutiles que en el caso de emociones actuadas, y por tanto, es más probable que dos anotadores humanos no escojan la misma categoría emocional para la misma elocución.

En este artículo, sugerimos la utilización de diversos coeficientes para estimar la fiabilidad de las anotaciones. En concreto, describimos cómo calcular la complejidad de la tarea de anotación mediante cálculos de entropía, así como el acuerdo entre anotadores empleando diferentes coeficientes Kappa. Las interpretaciones de los valores de estos coeficientes pueden llevar a error si se realizan siguiendo los enfoques tradicionales, que no tienen en cuenta las peculiaridades de las emociones no actuadas. Por ejemplo, se pueden obtener valores altos de entropía y bajas Kappa en corpus con aproximadamente el mismo número de elocuciones para cada emoción (incluyendo el neutro), puesto que es difícil discernir entre las emociones espontáneas. En cambio, en corpus adquiridos a partir de interacciones reales entre usuarios y un sistema de diálogo, como es nuestro caso, la categoría neutra es altamente predominante, lo que se traduce en valores bajos tanto de Kappa como de entropía. Por tanto, las interpretaciones tradicionales de estas medidas basadas en reglas estáticas no son fiables para ninguno de estos tipos de corpus, puesto que indicarían que las anotaciones no son fiables (alta entropía y Kappa bajo), o incluso producirían evaluaciones ambiguas (p.ej. baja Kappa y baja entropía).

Para obtener una interpretación fiable de los coeficientes, también sugerimos aportar diversas fuentes de información adicional. Concretamente, estudiar el acuerdo observado y las fuentes de desacuerdo tanto globalmente como entre cada par de anotadores, así como proveer valores contextuales que permitan interpretar los resultados obtenidos.

El resto del artículo esta estructurado de la siguiente forma. Las secciones 2 y 3 presentan respectivamente las medidas sugeridas para medir el acuerdo entre anotadores y la complejidad de la anotación de los corpus. La sección 4 propone un marco para la interpretación de los valores obtenidos. La sección 5 describe el corpus empleado en los experimentos junto con el proceso seguido para anotarlo con categorías emocionales, y presenta los resultados obtenidos empleando nuestra propuesta. Finalmente, la Sección 6 presenta las conclusiones.

2. *Medidas cuantitativas de acuerdo entre anotadores*

Tradicionalmente se ha abordado el estudio de los resultados de anotación empleando medidas que permiten evaluar el nivel de acuerdo entre los anotadores, de forma que con altas tasas de acuerdo se puede considerar que el resultado de una anotación es fiable. Con tal fin, se suelen calcular los denominados coeficientes Kappa ya que éstos tienen en cuenta hasta qué extremo los acuerdos observados no se deben al azar. Estos coeficientes se basan en la idea de prorratear la proporción de pares de anotadores que están de acuerdo (P_o) con la proporción de pares de anotadores que podrían estar de acuerdo tan sólo por azar (P_c). El resultado es la proporción entre el acuerdo observado no fortuito y todos los posibles acuerdos no fortuitos:

$$\kappa = \frac{P_o - P_c}{1 - P_c} \qquad (1)$$

El coeficiente Kappa más sencillo que podemos encontrar en la literatura es el propuesto por Fleiss (1971), al que hemos denominado multi-π siguiendo la notación de Artstein y Poesio (2005)[1]. El acuerdo observado de multi-π (P_o) se calcula como el número de casos en que dos anotadores distintos se ponen de acuerdo al etiquetar una elocución con la misma categoría emocional:

$$P_o = \frac{1}{UA(A-1)} \sum_{u=1}^{U} \sum_{e=1}^{E} n_{ue}(n_{ue} - 1) \qquad (2)$$

[1]Emplearemos esta notación en adelante para el resto de coeficientes Kappa.

En la Ecuación 2, n_{ue} representa el número de veces que la elocución 'u' ha sido anotada con la categoría emocional 'e'.

Fleiss asume que todos los anotadores comparten la misma distribución de probabilidad. Esto implica que la probabilidad de que un anotador clasifique una elocución 'u' con una emoción 'e' en particular, puede calcularse como la probabilidad global de anotar 'u' como 'e'. Esta probabilidad global se calcula como el número total de asignaciones a la emoción 'e' realizadas por todos los anotadores (n_e) dividida por el número de total de asignaciones ($U \cdot A$). El acuerdo fortuito (Ecuación 3) se ha calculado como la probabilidad de que cualquier par de anotadores etiqueten una misma elocución con la misma emoción, que hemos asumido como la probabilidad conjunta de que cada uno de ellos hiciese esta asignación de forma independiente, pues los anotadores juzgaron todas las elocuciones de forma independiente los unos de los otros.

$$P_c^{\pi} = \sum_{e=1}^{E} \left(\frac{1}{UA} n_e \right)^2 \tag{3}$$

El cálculo de multi-π asume que cada anotador sigue la misma distribución global de elocuciones en emociones. Sin embargo, dicha simplificación puede no ser plausible en cualquier dominio debido al efecto de las denominadas "tendencias de los anotadores" (*annotator bias*) en el valor de Kappa. En los experimentos realizados, dichas tendencias pueden definirse como la medida en que los anotadores difieren en la proporción de emociones dado un número particular de acuerdos. Con el resto de los parámetros fijados, el valor de Kappa crece al hacerlo dichas tendencias, esto es, cuando las proporciones de desacuerdos no son iguales para todas las emociones y hay un mayor desequilibrio entre ellas. Esta es la denominada *segunda paradoja del Kappa*. Podemos encontrar diferentes estudios de su impacto en la literatura, p.ej. (Feinstein y Cicchetti, 1990), (Cicchetti y Feinstein, 1990), (Lantz y Nebenzahl, 1996), y (Artstein y Poesio, 2005).

Para estudiar si la inclusión de distintos comportamientos de anotación mejora los valores de Kappa, se ha calculado la Kappa de Davies y Fleiss (1982), que se ha notado como multi-κ. El cálculo de multi-κ también se basa en la ecuación 1 y tiene el mismo acuerdo observado que multi-π (Ecuación 2). Sin embargo, para el acuerdo fortuito, incluye una distribución distinta para cada anotador.

Por tanto, en este caso la probabilidad de que un anotador 'a' clasifique una elocución 'u' con la emoción 'e' se calcula con el número observado de elocuciones asignadas a la emoción 'e' por ese anotador (n_{ae}), dividido por el número total de elocuciones (U). La probabilidad de que dos anotadores se pongan de acuerdo en anotar una elocución 'u' con la emoción 'e' es de nuevo la probabilidad conjunta de que cada uno realice dicha anotación de forma independiente:

$$P_c^{\kappa} = \frac{1}{\binom{A}{2}} \sum_{e=1}^{E} \sum_{j=1}^{A-1} \sum_{k=j+1}^{A} \frac{n_{aje}}{U} \frac{n_{ake}}{U} \tag{4}$$

A pesar de incluir diferencias entre los anotadores, multi-κ da a todos los desacuerdos la misma importancia. En la práctica, todos los desacuerdos no son igualmente probables y no tienen el mismo impacto sobre la calidad de los resultados de anotación. Por ejemplo, en nuestros experimentos, el desacuerdo entre *neutro* y *enfadado* es más fuerte que entre *neutro* y *dubitativo*, puesto que el primero se da entre dos categorías más fácilmente distinguibles.

Para tener en cuenta toda esta información, se han empleado coeficientes Kappa ponderados (Cohen, 1968)(Fleiss y Cohen, 1973), que se centran en los desacuerdos en lugar de en los acuerdos. Su cálculo se basa en la ecuación 5 (equivalente a la ecuación 1):

$$\kappa_w = 1 - \frac{\overline{P}_o}{\overline{P}_c} \tag{5}$$

donde $\overline{P}_o$ representa el desacuerdo observado y $\overline{P}_c$ el desacuerdo fortuito. Para todos los coeficientes empleados, el desacuerdo observado ha sido calculado como el número de veces que la elocución 'u' ha sido anotada con dos emociones distintas e_j y e_k por cada par de anotadores, ponderado según la distancia entre las emociones:

$$\overline{P}_o = \frac{1}{UA(A-1)} \sum_{u=1}^{U} \sum_{j=1}^{E-1} \sum_{k=j+1}^{E} X$$
$$X = n_{ue_j} n_{ue_k} distance(e_j, e_k) \tag{6}$$

Para calcular la distancia entre las emociones, proponemos situarlas en el espacio bidimensional denominado de *activación-evaluación*, en el que forman un patrón circular (Russell, 1980). Esto puede hacerse empelando disposiciones ya disponibles y ampliamente aceptadas en la literatura como la del estudio seminal (Plutchik, 1980), de forma que se pueda calcular la distancia angular entre las emociones.

Para optimizar los resultados, proponemos escoger siempre el ángulo menor entre las emociones consideradas (x o 360-x). De esta forma, la distancia entre cada dos ángulos estará siempre comprendida entre 0 y 180 grados. Para calcular los coeficientes, las distancias pueden expresarse como pesos con valores comprendidos entre 0 (0^o y por tanto ningún desacuerdo) y 1 (180^o y por tanto máximo desacuerdo).

En la literatura sobresalen los siguientes tres coeficientes Kappa ponderados: α, propuesto por Krippendorff (2003) y α' y β ambos propuestos por Artstein y Poesio (2005). Todos estos coeficientes comparten el mismo cálculo de desacuerdo observado (Ecuación 5). El desacuerdo por azar se calcula para α y α' de la siguiente forma:

$$\overline{P}_c^{\alpha} = \frac{1}{UA(UA-1)} \sum_{j=1}^{E-1} \sum_{k=j+1}^{E} X$$
$$X = n_{e_j} n_{e_k} distance(e_j, e_k) \quad (7)$$

$$\overline{P}_c^{\alpha'} = \frac{1}{(UA)^2} \sum_{j=1}^{E-1} \sum_{k=j+1}^{E} n_{e_j} n_{e_k} distance(e_j, e_k) \quad (8)$$

Como puede observarse en las Ecuaciones 7 y 8, estos coeficientes no consideran tendencias de los anotadores. Este problema puede resolverse empleando el coeficiente β con el que se tiene en cuenta el comportamiento observado de cada anotador:

$$\overline{P}_c^{\beta} = \sum_{j=1}^{E-1} \sum_{k=j+1}^{E} \left[\frac{1}{U^2 \binom{A}{2}} \sum_{m=1}^{A-1} \sum_{n=m+1}^{A} X \right]$$
$$X = n_{a_m e_j} n_{a_n e_k} distance(e_j, e_k) \quad (9)$$

3. *Medidas cuantitativas de la complejidad del proceso de anotación*

Si bien las emociones espontáneas proporcionan la mejor forma de estudiar las características de los comportamientos emotivos naturales, son difíciles de distinguir de los estados neutros. Esto hace que haya mayor probabilidad de que los anotadores de corpus constituidos por estas emociones anoten una misma elocución con distintas categorías emocionales. Por tanto, las tasas de acuerdo entre anotadores no deben ser la única fuente de información que se tome en consideración para evaluar la fiabilidad de estas anotaciones, pues la misma tasa de acuerdo en un corpus en el que las categorías sean fácilmente distinguibles indica un resultado mucho peor que en un corpus de emociones espontáneas, puesto que la complejidad de anotar el primero es mucho menor.

Diversos autores han empleado la entropía en el campo del procesamiento del lenguaje natural (Nesterenko y Rauzy, 2007). En su concepción tradicional, ésta fue definida en la teoría de la información de Shannon (1948) como la incertidumbre media de una única variable aleatoria, esto es, la cantidad de información derivada de esa variable. Trasladada al ámbito de la anotación de corpus afectivos, la entropía puede emplearse para medir la cantidad de información del proceso aleatorio que asigna a cada elocución una categoría emocional 'e' de entre un conjunto de E emociones.

Según esto, la entropía puede servir como medida cuantitativa de la complejidad de la tarea de anotación, puesto que si los anotadores llegan a muy altas tasas de acuerdo, habrá una entropía baja, lo que indicará que las categorías emocionales son claramente distinguibles. Por otra parte, si están en desacuerdo por lo general, se obtendrá un alto valor de entropía, lo que es un indicativo de lo dificultoso que puede llegar a ser para un anotador elegir una emoción que asignarle a cada elocución.

Siguiendo la propuesta de Steidl et al. (2005), la entropía puede calcularse como se muestra en la Ecuación 10:

$$H = \frac{1}{UA} \sum_{u=1}^{U} \sum_{a=1}^{A} H(a, u) \quad (10)$$

donde U es el número de elocuciones a anotar, A es el número de anotadores, y $H(a,u)$ es la entropía para la elocución 'u' y el anotador 'a'. $H(a,u)$ puede calcularse como sigue:

$$H(a,u) = -\sum_{e=1}^{E} l_e(\bar{a},u) Log_2(l_e(\bar{a},u)) \qquad (11)$$

donde E es el número de emociones, y $l_e(\bar{a},u)$ es la distribución de probabilidad de la variable aleatoria a la que hacíamos referencia anteriormente. Shannon quiso calcular con su teoría matemática el máximo teórico para la compresión de datos y su tasa de transmisión, que por lo general, se mide en bits, de ahí la presencia de Log_2. Puesto que $H(a,u)$ mide la cantidad de información del proceso aleatorio en que el anotador 'a' asigna una emoción a la elocución 'u', este no tiene por qué expresarse en base 2. Sin embargo, usar cualquier otra base no tendría más repercusión que un escalamiento linear de los resultados (Manning y Schütze, 2000).

Para considerar la incertidumbre introducir por el anotador 'a', $l_e(\bar{a},u)$ se calcula siguiendo la Ecuación 12, donde l_{ref} no considera al anotador 'a' ($\bar{a}$), y l_{dec} considera a todos los anotadores (Steidl et al., 2005). Ambas medidas representan el número de veces que la categoría emocional 'e' ha sido escogida para la elocución 'u', promediada por el número total de anotadores. Es decir, estas medidas calculan el comportamiento observado del proceso aleatorio descrito anteriormente.

$$l_e(\bar{a},u) = \frac{l_{ref_e}(\bar{a},u) + l_{dec_e}(u)}{2} \qquad (12)$$

4. *Propuestas para la interpretación de la fiabilidad*

Cuando se compilan corpus a partir de interacciones orales espontáneas, la mayoría de las elocuciones se corresponden con un estado neutro del usuario. Esto provoca que, incluso con una alta tasa de acuerdo entre anotadores, el valor de los coeficientes Kappa sea bajo. La situación en la que a pesar de tener un número prácticamente idéntico de acuerdos, la distribución de éstos entre las categorías de anotación afecta profundamente a los coeficientes Kappa se conoce como la *primera paradoja del Kappa*. Este fenómeno establece que, siendo el resto de parámetros iguales, el valor de Kappa crece con distribuciones simétricas de los acuerdos. Es decir, si una categoría predomina claramente sobre las demás, entonces el acuerdo fortuito (P_c) es alto y la Kappa decrece considerablemente (Feinstein y Cicchetti, 1990)(Cicchetti y Feinstein, 1990). Por consiguiente, la primera paradoja del Kappa puede afectar enormemente a los valores de los coeficientes y por tanto debe ser tenida en cuenta en su interpretación.

Tradicionalmente, las interpretaciones de Kappa están basadas en reglas estáticas, siendo una de las más globalmente empleadas la propuesta por Landis y Koch (1977), en la que se realiza una correspondencia entre intervalos de valores Kappa e interpretaciones de los acuerdos. Alternativamente, Krippendorff (2003) estableció 0,65 como umbral para la aceptabilidad de los resultados de acuerdo. Sin embargo, emplear intervalos o umbrales pre-establecidos de valores Kappa no ofrece la suficiente información como para realizar una interpretación justificada de la aceptabilidad de los resultados.

Por otra parte, en los trabajos en la literatura en que se emplea la entropía como medida de la complejidad de la anotación, se trata con corpus que han sido confeccionados para que el número de elocuciones en cada categoría sea similar, p.ej. (Steidl et al., 2005). Este tipo de corpus es interesante para estudiar las características de emociones no actuadas y comparar los resultados de clasificación automática y humana de emociones expresadas de forma sutil. Sin embargo, estos corpus no responden al desequilibrio intrínseco de una interacción real en la que el número de elocuciones que reflejan un comportamiento emocional es mucho menor que aquellas que responden a un estado neutro. Por consiguiente, el cálculo de la entropía conduciría inevitablemente a interpretaciones positivas, puesto que es muy fácil para los anotadores coincidir en la categoría neutra. Por tanto, la interpretación de la entropía sería justamente contraria a la interpretación de los coeficientes Kappa, lo que podría llevar a malentendidos.

Para resolver estos problemas, proponemos aportar información adicional de dos modos: i) estudiando los acuerdos entre anotadores así como las fuentes principales de desacuerdo y ii) encuadrando los coeficientes

Kappa y la entropía en contexto para alcanzar una interpretación que conjugue el máximo número de factores posible y que por tanto, permita valorar la adecuación de las interpretaciones tradicionales.

Puesto que los coeficientes Kappa tienen en cuenta el acuerdo fortuito, suelen preferirse al cálculo de las tasas de acuerdo observado, información que pocas veces se aporta como parte del proceso de interpretación de la fiabilidad. Sin embargo, debido a las dificultades relativas a la interpretación de los coeficientes Kappa en corpus de emociones no actuadas, proporcionar el acuerdo observado es muy valioso. Por tanto, en primer lugar proponemos incluir no sólo el acuerdo observado global, sino también el acuerdo por pares (tasa de acuerdo entre cada par de anotadores para todas las elocuciones). Asimismo, también es recomendable calcular los valores de entropía para cada anotador. Para hacerlo, proponemos emplear la Ecuación 11, promediando por el número de elocuciones y anotadores considerados sin excluir el anotador que se estudie en cada caso.

En segundo lugar, para aportar información suficiente para hacer una interpretación justificada de fiabilidad, proponemos emplazar los coeficientes Kappa en contexto calculando sus valores máximo, mínimo y normal a partir del acuerdo observado, tal y como se indica en (Lantz y Nebenzahl, 1996). Para el mismo acuerdo observado, los valores de Kappa pueden variar considerablemente entre $kappa_{min}$ y $kappa_{max}$ dependiendo del equilibrio del corpus. El valor $Kappa_{max}$ se obtiene cuando se desequilibran al máximo los desacuerdos al mismo tiempo que se mantienen equilibrados los acuerdos, mientras que $kappa_{min}$ se obtiene cuando los acuerdos están desequilibrados y los desacuerdos equilibrados. $Kappa_{nor}$ no corresponde a un valor ideal de Kappa, sino a distribuciones simétricas tanto de acuerdos como de desacuerdos. Como se describe en (Lantz y Nebenzahl, 1996), los desplazamientos respecto al valor $kappa_{nor}$ indican asimetría en los acuerdos o desacuerdos dependiendo de si están más cerca del valor mínimo o máximo respectivamente. Por tanto, aportar estos valores ayuda a comprender mejor la complejidad de la tarea debido al desequilibrio inherente a estos corpus, lo que adquiere especial relevancia al justificar las interpretaciones de los valores obtenidos.

El caso de la entropía es distinto en cuanto que ésta siempre tiene el mismo valor mínimo (cero) en el caso en que todos los anotadores estén de acuerdo. El valor máximo de $H(a, u)$ responde a una distribución equiprobable de los valores de la variable aleatoria. En nuestro caso, esto se traduce al supuesto en que cada anotador asigna cualquier categoría con igual probabilidad. Para maximizar H, la emoción escogida de forma equiprobable debería ser distinta para cada anotador en cada elocución, lo que no es factible en la práctica cuando el número de anotadores es distinto al de emociones. Por tanto, no es trivial calcular el valor máximo si durante el proceso de anotación no hubo tantas categorías como anotadores. En cualquier caso, se puede aportar un cálculo aproximado de la entropía máxima que arroje luz sobre la interpretación del valor de dicho coeficiente.

5. Resultados experimentales

El sistema de diálogo UAH (Universidad al Habla) fue desarrollado en nuestro laboratorio para proporcionar información académica telefónicamente, de forma oral. El corpus empleado para los experimentos descritos en este artículo, está compuesto por 85 diálogos de 60 usuarios distintos interactuando con el sistema (Callejas y López-Cózar, 2008). El corpus contiene 422 turnos de usuario y tiene una duración de 150 minutos. Nueve anotadores clasificaron cada elocución del corpus con una de las siguientes emociones: enfadado, aburrido, dubitativo y neutro. La categoría asignada finalmente a cada elocución se decidió por mayoría. De media, más del 80 % de las elocuciones se anotaron como neutras.

Para evaluar la fiabilidad de la anotación del corpus, calculamos los coeficientes tal y como se describe en las Secciones 2 y 3. Para los coeficientes Kappa, obtuvimos los resultados que se muestran en la Figura 1. Como puede observarse, una interpretación tradicional de estos valores indicaría un acuerdo tan sólo aceptable, o incluso inaceptable, siguiendo las interpretaciones tradicionales.

Para la entropía, sin embargo, el valor obtenido es 0,318, que por el contrario indica alta fiabilidad en la notación. Para aportar suficiente información para desambiguar las interpretaciones contradictorias obtenidas, seguimos el método propuesto en la Sección 4.

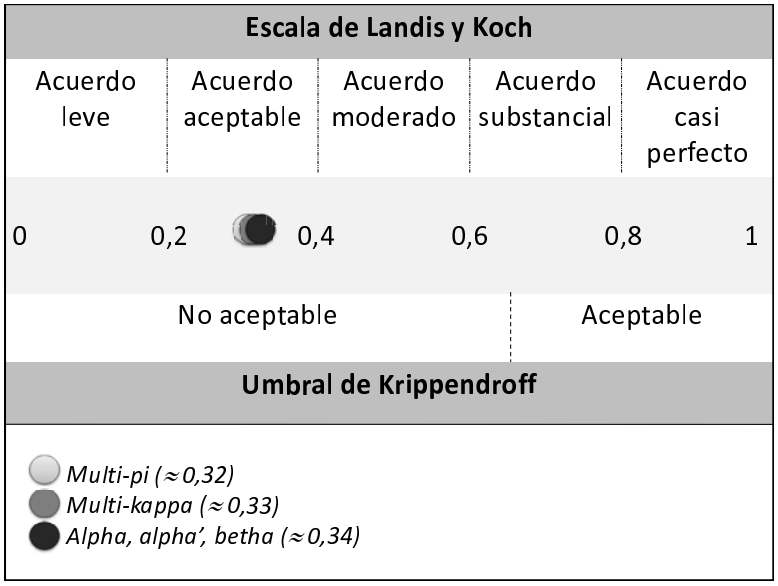

Figura 1: Valores de los coeficientes Kappa y sus interpretaciones tradicionales

5.1. Cálculo del desacuerdo entre anotadores

En primer lugar, calculamos el acuerdo observado entre los anotadores y encontramos que era mayor de 0,85. Es interesante destacar que la gran diferencia entre este valor y los coeficientes Kappa (que están todos alrededor de 0,32) se debe a la alta probabilidad de acuerdo fortuito. El alto acuerdo fortuito y la baja entropía indican que fue fácil para los anotadores coincidir en la misma emoción, lo que muestra que hay un importante desequilibrio entre las categorías del corpus, lo que a su vez explica los bajos valores Kappa obtenidos. En otras palabras, estos valores tan bajos no se deben a falta de fiabilidad en la anotación, sino a la alta tasa de neutros íntrinseca a los corpus de emociones espontáneas.

Además, como se observa en la Figura 1, ponderar los desacuerdos (β y α vs. multi-κ) reduce los valores de Kappa, lo que significa que las principales fuentes de desacuerdo tienen lugar para las categorías más distantes. A esta misma conclusión se llegó tras añadir información acerca de los acuerdos por pares, tal y como sugerimos en la Sección 4. De esta manera, corroboramos que no había muchos desacuerdos (corroborado por la alta tasa de acuerdo observado). Además los desacuerdos existentes ocurrían en la mayoría de los casos entre categorías neutras y no neutras, que además son las más distantes. En cambio, hubo pocos desacuerdos entre categorías no neutras. Esto es otra consecuencia de emplear emociones espontáneas, ya que debido a su sutileza, son difíciles de distinguir de los estados neutros.

Por otra parte, la entropía calculada para cada anotador también indica un buen acuerdo por pares, ya que los valores tan sólo variaron entre 0,317 y 0,320, con una media de 0,318.

5.2. Contextualización de las medidas de evaluación

En segundo lugar, como propusimos en la Sección 4, calculamos el contexto para todos los coeficientes. En el caso de los coeficientes Kappa (Figura 2), nuestros resultados señalan que aportar los valores de Kappa es más informativo cuando están en contexto pues obtenemos información muy valiosa acerca de posibles desequilibrios, necesaria por otra parte para llegar a conclusiones apropiadas. Por ejemplo, en nuestro caso se dieron desplazamientos considerables del valor $Kappa_{nor}$ en todos los casos, lo que corrobora que hay una gran asimetría entre categorías. Esto se debe una vez más al fenómeno de prevalencia discutido anteriormente (primera paradoja Kappa) y no a una anotación incorrecta del corpus.

En el caso de la entropía, el valor máximo se da en caso de equiprobabilidad de emociones y maximo desacuerdo entre los anotadores para la misma elocución, es decir, cuan-

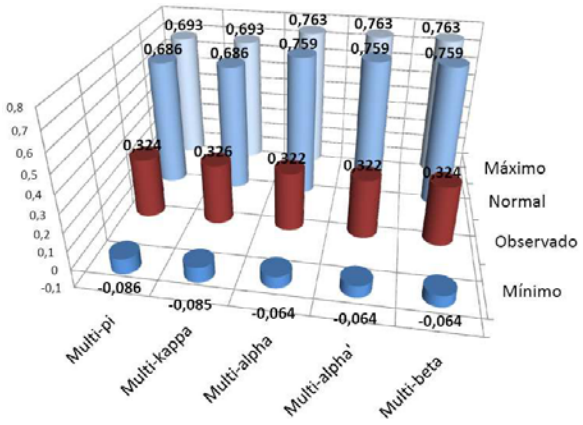

Figura 2: Valores mínimo, máximo, normal y observado de los coeficientes Kappa

do cada categoría se elige un mínimo número de veces para cada elocución (desacuerdo total). En nuestro caso, el mínimo número de repeticiones es 2 para 3 de las categorías (la categoría emocional es escogida por dos anotadores) y 3 para una categoría (puesto que consideramos un número impar de anotadores). El valor de la entropía obtenido fue 0,988, que está muy lejos del 0,318 observado en nuestro corpus. Cuando calculamos la entropía por anotador, los valores fluctuaban entre 0,987 y 0,990, cifras mucho mayores que la entropía observada. Estos resultados se deben también al predominio inevitable de la categoría 'neutro'.

6. Conclusiones

En este artículo hemos propuesto el uso de diversos coeficientes y fuentes de información para mejorar la evaluación de la anotación humana de corpus afectivos. Nuestra propuesta ha sido evaluada empíricamente con un corpus de emociones no actuadas recogidas a partir de llamadas espontáneas a un sistema de diálogo. Los resultados experimentales subrayan la dificultad de evaluar la fiabilidad de este tipo de corpus. Las interpretaciones tradicionales del valor de coeficientes que miden el acuerdo entre anotadores y la complejidad de la tarea, consideraría la anotación de estos corpus no fiable. Sin embargo, esta conclusión no sería adecuada puesto que estas interpretaciones no tienen en cuenta los efectos del desequilibrio inherente a estos corpus, ni la sutileza de emociones espontáneas. Nuestro método mejora la interpretación de los resultados tradicionales y provee de una base sólida para evaluar con más información la calidad de las anotaciones.

Bibliografía

Artstein, Ron y Massimo Poesio. 2005. *kappa*$_3$ = alpha (or beta). Informe técnico, University of Essex.

Callejas, Zoraida y Ramón López-Cózar. 2008. Relations between de-facto criteria in the evaluation of a spoken dialogue system. *Speech Communication*, 50(8-9):646–665.

Cicchetti, Domenic V. y Alvan R. Feinstein. 1990. High agreement but low Kappa: II. Resolving the paradoxes. *Journal of Clinical Epidemiology*, 43(6):551–558.

Cohen, Jacob. 1968. Weighted kappa: nominal scale agreement with provision for scaled disagreement or partial credit. *Psychological Bulletin*, 70(4):213–220.

Davies, Mark y Joseph L. Fleiss. 1982. Measuring agreement for multinomial data. *Biometrics*, 38(4):1047–1051.

Douglas-Cowie, Ellen, Nick Campbell, Roddy Cowie, y Peter Roach. 2003. Emotional speech: towards a new generation of databases. *Speech Communication*, 40:33–60.

Feinstein, Alvan R. y Domenic V. Cicchetti. 1990. High agreement but low Kappa: I. The problems of two paradoxes. *Journal of Clinical Epidemiology*, 43(6):543–549.

Fleiss, Joseph L. 1971. Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76(5):378–382.

Fleiss, Joseph L. y Jacob Cohen. 1973. The equivalence of weighted kappa and the interclass correlation coefficient as measures of reliability. *Educational and Psychological Measurement*, 33:613–619.

Johnstone, T. 1996. Emotional speech elicited using computer games. En *Proceedings of the 4th International Conference on Spoken Language Processing (ICSLP 1996)*, volumen 3, páginas 1985–1988, Philadelphia, PA.

Krippendorff, Klaus. 2003. *Content Analysis: An Introduction to its Methodology*. Sage Publications, Inc.

Landis, J. R. y G. G. Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics*, 33:159–174.

Lantz, Charles A. y Elliott Nebenzahl. 1996. Behavior and interpretation of the κ statistic: Resolution of the two paradoxes. *Journal of Clinical Epidemiology*, 49(4):431–434.

Manning, Christopher D. y Hinrich Schütze. 2000. *Foundations of statistical natural language processing*. The MIT Press.

Nesterenko, Irina y Stéphane Rauzy. 2007. On the use of probabilistic grammars in speech annotation and segmentation tasks. En *Proceedings of SPECOM 2007*, Moscú, Rusia.

Plutchik, Robert. 1980. *EMOTION: A psychoevolutionary synthesis*. Harper and Row publishers.

Russell, J A. 1980. A circumplex model of affect. *Journal of Personality and Social Psychology*, 39:1161–1178.

Shannon, C. E. 1948. A mathematical theory of communication. *Bell System Technical Journal*, 27:379–423.

Steidl, Stefan, Michael Levit, Anton Batliner, Elmar Nöth, y Heinrich Niemann. 2005. Of all things the measure is man. automatic classification of emotions and inter-labeler consistency. En *Proceedings of ICASSP 2005*, páginas 317–320, Philadelphia, USA.

Vogt, Thurid y Elisabeth André. 2005. Comparing feature sets for acted and spontaneous speech in view of automatic emotion recognition. En *Proceedings of IEEE International Conference on Multimedia and Expo*, páginas 474–477.

Wilting, Janneke, Emiel Krahmer, y Marc Swerts. 2006. Real vs. acted emotional speech. En *Proceedings of Interspeech 2006*, páginas 805–808, Pittsburgh PA, USA.