EVALUACIÓN DE CAMPO DE UN SISTEMA DE DIÁLOGO ORAL EMPLEANDO RELACIONES ESTADÍSTICAS

Zoraida Callejas, Ramón López-Cózar

Departamento de Lenguajes y Sistemas Informáticos
Universidad de Granada. 18071 Granada
{zcallejas,rlopezc}@ugr.es

RESUMEN

La evaluación de sistemas de diálogo oral se puede realizar mediante medidas "objetivas" calculadas de forma instrumental o mediante expertos y con juicios de opinión de los usuarios que hayan empleado el sistema con anterioridad (medidas "subjetivas"). En la literatura podemos encontrar diversos trabajos que tratan de establecer relaciones entre ambos tipos de medidas. En este artículo describimos los resultados empíricos obtenidos de estudios estadísticos sobre interacciones de usuarios reales con un sistema de diálogo experimental. Estos estudios no han sido suficientemente explorados en la literatura y como demostramos, pueden mostrar relaciones importantes entre criterios de evaluación, que pueden servir de guía para refinar los sistemas que se evalúan, así como para contribuir al conocimiento acerca de cómo los aspectos cuantitativos pueden afectar la percepción del usuario acerca del sistema.

1. INTRODUCCIÓN

Con el fin de minimizar costes y optimizar resultados, existe la necesidad de encontrar métodos, arquitecturas y criterios estándar para evaluar, comparar y predecir el rendimiento y la usabilidad de los sistemas de diálogo. Numerosas investigaciones realizadas durante los años 90 (p.e. [1][2]) sentaron las bases para establecer un conjunto común de criterios cuantitativos de evaluación. Sin embargo, no existe un consenso global acerca de qué criterios deben tenerse en cuenta para optimizar la usabilidad de los sistemas de diálogo. Algunos proyectos han intentado abordar el problema de la predicción de la usabilidad y la satisfacción del usuario a partir de criterios de rendimiento medibles. Este es el caso del modelo PARADISE [3], que se ha convertido en uno de los modelos de referencia para la evaluación de este tipo de sistemas.

Debido a la complejidad y el esfuerzo que demanda la aplicación de este modelo, muchos autores aplican medidas cualitativas y cuantitativas por separado. Por ejemplo, el sistema multimodal de navegación MUMS [4], el sistema VIrtual CO-driver [5], el quiosco multimedia MASK [6] y el sistema de diálogo SAMMIE [7], se han evaluado únicamente de forma subjetiva. Otros autores, p.e. [8], evalúan sus sistemas tanto con criterios medidos instrumentalmente como con opiniones de los usuarios acerca de su calidad, pero sin establecer enlaces entre las distintas medidas de evaluación empleadas.

Por otra parte, los resultados en la literatura están por lo general basados en interacciones restringidas en laboratorio, en las que se solicita a los usuarios que interactúen con el sistema siguiendo unos escenarios previamente establecidos. La principal desventaja de este método es que dichos escenarios pueden diferir de las tareas que un usuario habría seleccionado en una interacción no predefinida. Por el contrario, la evaluación de campo se realiza a partir de interacciones de usuarios reales con el sistema final en sus entornos reales. Los resultados obtenidos mediante evaluaciones de campo son robustos ante la heterogeneidad de usuarios, dispositivos y entornos; por consiguiente, son más relevantes para la predicción del comportamiento real de los sistemas que los estudios de laboratorio.

La contribución del artículo al estado del arte en la evaluación de sistemas de diálogo oral consiste en la obtención de nuevas evidencias empíricas por medio de un estudio de campo llevado a cabo sobre nuestro sistema de diálogo experimental UAH (Universidad al Habla). Para ello se han empleado estudios de correlación, estableciendo relaciones entre criterios tanto cuantitativos como cualitativos.

2. CRITERIOS DE EVALUACIÓN

Este artículo presenta los resultados de evaluación del sistema UAH, que se desarrolló para proveer acceso telefónico automático a información académica de nuestro Departamento [9]. La evaluación del sistema se ha llevado a cabo tanto con parámetros de interacción como con juicios de calidad. Los primeros han sido extraídos de forma semi-automática a partir de los diálogos grabados, mientras que los segundos se han obtenido de cuestionarios que los usuarios podían rellenar de forma voluntaria.

Para calcular los parámetros de interacción, hemos empleado un corpus de diálogos construido a partir de las llamadas telefónicas realizadas al sistema UAH por alum-

Este trabajo ha sido subvencionado por el proyecto HADA TIN2007-64718 (Ministerio de Educación y Ciencia).

nos de nuestra Universidad durante su primer año de utilización. Este corpus consta de 85 diálogos y 422 turnos de usuario, con una media de 5 turnos por diálogo. Los parámetros de evaluación empleados han sido los siguientes: éxito de la tarea, completitud del diálogo, duración del diálogo, número de turnos de usuario, media de palabras por turno, WER, confianza de reconocimiento media, porcentaje de elocuciones correctamente comprendidas, y número de turnos de confirmación.

Las medidas de evaluación subjetivas que hemos empleado han sido las siguientes: percepción de hasta qué punto UAH entiende al usuario, percepción de hasta qué punto el usuario entiende a UAH, velocidad de interacción percibida, presencia percibida de errores cometidos por UAH, facilidad percibida de corregir los errores de UAH, facilidad percibida de conseguir la información requerida, satisfacción del usuario, percepción de hasta qué punto el usuario sabía qué hacer en cada momento de la interacción y percepción de hasta qué punto UAH se comportaba de forma similar a un ser humano. Además, también se ha extraído de los cuestionarios el nivel de conocimiento técnico de los usuarios y su experiencia previa utilizando el sistema.

3. ESTUDIOS ESTADÍSTICOS

Para encontrar relaciones relevantes entre los criterios utilizados, hemos correlacionado todas las variables, obteniendo el valor absoluto del *coeficiente de correlación de Pearson*, así como la significatividad (o *p-value*) de cada coeficiente de correlación.

Dado que la mayoría de las variables estaban intercorrelacionadas, se ha estudiado el efecto que cada criterio ejercía en la significatividad de las relaciones entre los demás criterios. Para estudiar las relaciones aisladamente, eliminando el efecto del resto de los criterios, hemos medido los *coeficientes de correlación parcial* conjuntamente con sus niveles de significación.

El *coeficiente de correlación de Pearson* funciona correctamente para las variables escalables (p.e. la duración del diálogo), pues es apropiado realizar comparaciones de distancia entre valores. Sin embargo, para la investigación descrita también se han empleado variables ordinales (p.e. parámetros de calidad percibida) y dicotómicas (p.e. "éxito de la tarea" o "completitud del diálogo"). Para obtener resultados fiables, se han generado tablas de contingencia para los criterios ordinales así como los coeficientes *Tau-b de Kendall* y *Rho de Spearman*.

Además, hemos llevado a cabo análisis de varianza utilizando el test *one-way ANOVA* junto con el *coeficiente F*. Para obtener más información en la que basar las interpretaciones realizadas, y especialmente para el caso de las variables dicotómicas, también se ha calculado el valor de la *V de Cramer*, que permite contrastar la hipótesis de independencia en las tablas de contingencia.

4. DISCUSIÓN DE LOS RESULTADOS

Los dos valores más altos de correlación con la satisfacción del usuario han sido obtenidos en todos los estudios estadísticos para los criterios siguientes: "facilidad percibida de conseguir la información requerida" y "éxito de la tarea". Según lo esperado, la satisfacción del usuario es alta cuando éste consigue fácilmente la información que requiere. Sin embargo, cabe destacar que el modo en que los usuarios obtienen la información tiene respecto a su satisfacción, la misma significatividad que el hecho de que finalmente consigan dicha información. En [10], la satisfacción del usuario también está correlacionada con que el usuario obtuviera finalmente la información que buscaba. Sin embargo, el indicador de Möller de facilidad de la comunicación no proporcionaba una contribución significativa a la satisfacción total del usuario. Este hecho puede sugerir que la facilidad de la comunicación es más importante para los usuarios que tienen una necesidad verdadera de obtener la información (estudios de campo), que para quienes la interacción se realiza siguiendo escenarios predefinidos (estudios de laboratorio).

Además, Rajman et al. [11] mostraron que, dado que los usuarios en evaluaciones de laboratorio no tienen la posibilidad de contrastar la información proporcionada por el sistema de diálogo, éstos confían ciegamente en las respuestas del sistema. Es decir, no comprueban si la información es correcta o útil, y por tanto, consideran el hecho de obtener una respuesta del sistema equivalente a obtener un resultado correcto. En nuestros experimentos, se ha proporcionado a los usuarios información académica real. Dado que necesitaban realmente esta información, podían contrastarla y saber si era exacta o no. Así, entre los diálogos no exitosos (tanto desde el punto de vista de los parámetros de la interacción como de las valoraciones sobre la calidad) se han dado casos donde a pesar de que el sistema proporcionó al usuario información, ésta no era la que él deseaba, como demuestra el hecho de que algunos diálogos completos no fueron exitosos. La evaluación de campo presenta, de este modo, la gran ventaja de posibilitar una separación entre la calidad de la interacción y la calidad de los resultados obtenidos.

Centrándonos en los parámetros de la interacción, hay una correlación notable entre la completitud del diálogo y el éxito de la tarea. Aunque los usuarios podían finalizar la llamada en cuanto recibían la información deseada, éstos esperaron generalmente hasta el final en los diálogos exitosos. Este hecho difiere de los resultados de otros autores. Por ejemplo, Turunen et al. [12] mostraron que había diferencias significativas entre la forma de llevar a cabo la interacción en las pruebas de laboratorio y en evaluaciones de campo con el sistema Stopman. En su evaluación de campo menos de un 10 % de los usuarios esperaron al final de la llamada antes de colgar. El número de diálogos en los cuales los usuarios esperaron hasta el final de la interacción (es decir, el número de diálogos completos) en nuestro estudio de campo es un 50 % mayor que el mos-

trado en [12], seguramente debido a una actitud "tecnofílica" de nuestros usuarios, en su mayoría estudiantes de la Escuela de Ingeniería Informática y Telecomunicaciones.

Otro criterio que está estrechamente correlacionado con el éxito de la tarea y la satisfacción del usuario es la facilidad percibida para corregir errores. Sin embargo, la presencia percibida de errores no se correla con ninguno de estos criterios. Esto puede deberse a que, aunque en el 48,19 % de los diálogos exitosos los usuarios han detectado errores, en la mayoría de los casos han sabido corregirlos y obtener la información que buscaban. Concretamente, un 69,23 % de los usuarios han considerado "fácil" o "muy fácil" corregir errores en los diálogos exitosos. Sin embargo, en los no exitosos, un 83,33 % de los usuarios ha manifestado que la corrección de errores era "difícil" o "muy difícil".

En [10], la opinión de los usuarios sobre si los malentendidos podrían ser aclarados fácilmente (que se clasificó como un factor que contribuía a la calidad del diálogo), no resultó ser un buen indicador de la satisfacción del usuario. Además, el autor encontró que la satisfacción del usuario no se podía predecir completamente mediante el éxito de la tarea, y sostuvo que este resultado podría ser debido a las condiciones poco realistas de la experimentación de laboratorio empleada en su investigación. Por tanto, se ha corroborado este hecho en nuestro estudio de campo, puesto que los cuestionarios subjetivos no se han podido substituir por los parámetros de la interacción empleados sin que esto supusiera pérdida de información.

Por otra parte, el criterio que ha mostrado un mayor número de correlaciones significativas ha sido la "percepción de hasta qué punto UAH entiende al usuario". Las relaciones más significativas entre esta valoración de la calidad y otros parámetros han sido obtenidos con el éxito de la tarea y la satisfacción del usuario. Cabe también destacar que el grado con el cual el usuario percibe que el sistema UAH le entiende no está correlacionado con los parámetros de la interacción que miden el funcionamiento del reconocedor del habla, como WER o medidas de confianza. Sin embargo, sí está correlacionado con el porcentaje de elocuciones correctamente entendidas, ello indica que desde el punto de vista del usuario, los errores de reconocimiento del habla no son importantes siempre y cuando las interpretaciones semánticas sean correctas y estos errores sean imperceptibles para el usuario.

4.1. Influencia de la iniciativa de gestión del diálogo

Para estudiar la influencia de la iniciativa utilizada para la gestión del diálogo, hemos repetido la experimentación comentada anteriormente, pero distinguiendo entre los diálogos con iniciativa dirigida por el sistema y los diálogos con iniciativa mixta.

El éxito de la tarea es aproximadamente igual para ambas iniciativas de gestión del diálogo. Este resultado difiere de los que se pueden encontrar en la literatura, p.e. [10], donde una iniciativa más flexible conduce a tasas de éxito considerablemente más altas. En nuestros experimentos el éxito es mayor para la iniciativa mixta, pero la diferencia entre ambas es insignificante pues el 77,77 % de los diálogos de iniciativa mixta y el 76,92 % de los diálogos con iniciativa por parte del sistema han concluido con éxito.

Sin embargo, a la luz de los resultados experimentales, el éxito de la tarea parece estar relacionado con distintos factores en cada tipo de iniciativa. De esta manera, en la iniciativa mixta la seguridad del usuario sobre qué hacer en cada momento del diálogo no está correlacionada con el éxito de la tarea, la satisfacción del usuario ni la percepción sobre la facilidad de obtener la información requerida. Por el contrario, el éxito de la tarea tiene una correlación significativa con la seguridad del usuario en los diálogos dirigidos por el sistema. Este hecho sucede probablemente porque el usuario dispone de mayor libertad en las interacciones con iniciativa mixta y, por tanto, el sistema no restringe lo que debe decir en cada momento. Sin embargo, esta situación no conduce a malos resultados de la interacción, pues el éxito de la tarea no se reduce al emplear iniciativa mixta.

Las correlaciones de la facilidad percibida de conseguir la información requerida son también muy diferentes en ambos casos. En el caso de la iniciativa por parte del sistema, está relacionada con la completitud del diálogo, el porcentaje de elocuciones correctamente comprendidas y la opinión que el usuario tiene sobre el comportamiento humano del sistema. Por el contrario, para los diálogos con iniciativa mixta, la facilidad percibida no está correlacionada con estas medidas, sino con indicadores de la duración de la interacción como la "duración del diálogo" o el "número de turnos de usuario". Igualmente sucede con la satisfacción y el éxito de la tarea, que están altamente correlacionadas con medidas de duración en interacciones con iniciativa mixta, pero no en los diálogos dirigidos por el sistema. La duración de estos diálogos está correlacionada de manera perceptible con la satisfacción del usuario, mientras que en sistemas con iniciativas más estrictas de la interacción, no es considerada tan importante por los usuarios. Además, la duración media de los diálogos es menor cuando la interacción es más flexible.

Los estudios basados en pruebas de laboratorio como los de Rajman et al. [11] no han podido percibir variaciones claras en la calidad con respecto al predominio de la iniciativa del sistema o del usuario. Además, algunas pruebas de laboratorio como las llevadas a cabo en [10] para el sistema BoRIS no pudieron encontrar ninguna relación significativa entre la iniciativa y otros parámetros de la interacción. Sin embargo, nuestros resultados demuestran que la significación de las relaciones entre los diversos criterios de evaluación, incluyendo parámetros de la interacción y valoraciones de la calidad, varía dependiendo de la iniciativa utilizada para la gestión del diálogo.

5. CONCLUSIONES

En este artículo se ha presentado un estudio de las relaciones entre varios criterios estándar de-facto para la evaluación de un sistema de diálogo oral con el que se interactúa telefónicamente. Nuestros resultados experimentales se basan en un estudio de campo que utiliza interacciones reales registradas por usuarios no reclutados previamente que han llamado espontáneamente al sistema para obtener información.

Para realizar nuestro estudio se han calculado parámetros de la interacción (o medidas objetivas) y juicios de la calidad (medidas subjetivas) empleando un corpus de las interacciones reales sistema-usuario. Se han llevado a cabo un conjunto de estudios estadísticos a partir de los cuales se han extraído relaciones significativas entre todos los criterios.

Nuestros resultados demuestran que el éxito de la tarea, la facilidad percibida de obtener la información y el punto hasta el cual el usuario percibe que el sistema le entiende están correlacionados con la satisfacción del usuario. Estos resultados sugieren que obtener la información requerida no conlleva necesariamente la satisfacción del usuario, dado que los usuarios valoraron en algunos casos que diálogos exitosos no les habían satisfecho debido a que encontraron dificultades para obtener la información que estaban buscando (a pesar de haber recibido datos que concordaban con su petición). Ésta es una de las implicaciones derivadas del uso de los estudios de campo, en los cuales los usuarios se preocupan no sólo de obtener la información que buscaban, sino también de obtenerla fácilmente y de que ésta sea correcta.

Además, la relación entre la facilidad percibida de obtener la información y otros criterios varía notablemente con la estrategia de gestión de diálogo empleada. Los datos estadísticos sugieren que la predicción de la satisfacción del usuario también depende de la iniciativa del diálogo empleada. En los diálogos con iniciativa mixta parece estar relacionada más directamente con medidas objetivas, como la duración del diálogo. Sin embargo, en diálogos más restringidos, las medidas subjetivas como el grado hasta el cual el usuario percibe que el sistema le entiende, tienen un impacto mayor. Se trata de un resultado importante que podría indicar una necesidad de adaptar los procedimientos de evaluación al tipo de interacciones que se analizan.

6. BIBLIOGRAFÍA

[1] M.A. Walker, A. Rudnicky, R. Prasad, J. Aberdeen, E. Owen Bratt, J. Garofolo, H. Hastie, A. Le, B. Pellom, A. Potamianos, R. Passonneau, S. Roukos, G. Sanders, S. Seneff, y D. Stallard, "DARPA Communicator: Cross-System Results for the 2001 Evaluation," in *Proc. of ICSLP'02*, Denver, USA, 2002, vol. 1, pp. 269–272.

[2] EAGLES, "Evaluation of Natural Language Processing Systems. Final report. Document EAG-EWG-PR2," Tech. Rep., Center for Sprogetknologi, Copenhague, Denmark, 1996.

[3] M. Walker, C. A. Kamm, y D. J. Litman, "Towards developing general models of usability with PARADISE," *Natural Language Engineering*, pp. 363–377, 2000.

[4] T. Hurtig, "Visualization and multimodality: a mobile multimodal dialogue system for public transportation navigation evaluated," in *Proc. of MobileHCI'06*, Helsinki, Finland, 2004, pp. 251–254.

[5] P. Geutner, F. Steffens, y D. Manstetten, "Design of the VICO Spoken Dialogue System: Evaluation of User Expectations by Wizard-of-Oz experiments," in *Proc. of LREC'02*, Las Palmas de Gran Canaria, Spain, 2002.

[6] L. Lamel, S. Bennacef, J.L. Gauvain, H. Dartigues, y J.N. Temem, "User evaluation of the MASK kiosk," *Speech Communication*, vol. 38, no. 1-2, pp. 131–139, 2002.

[7] T. Becker, C. Gerstenberger, I. Kruijff-Korbayova, A. Korthauer, M. Pinkal, M. Pitz, P. Poller, y J. Schehl, "Natural and intuitive multimodal dialogue for In-Car Applications: The SAMMIE System," in *Proc. of PAIS'06*, Riva del Garda, Italy, 2006, pp. 612–616.

[8] S. M. Robinson, A. Roque, As. Vaswani, y D. Traum, "Evaluation of a spoken dialogue system for virtual reality call for fire training," in *Proc. of the 25th Army Science Conference*, Orlando, USA, 2006.

[9] Z. Callejas y R. López-Cózar, "Implementing modular dialogue systems: a case study," in *Proc. of ASIDE'05*, Aalborg, Denmark, 2005.

[10] S. Möller, *Quality of telephone-based spoken dialogue systems*, Springer, 2005.

[11] M. Rajman, T. H. Bui, A. Rajman, F. Seydoux, A. Trutnev, y S. Quarteroni, "Assessing the usability of a dialogue management system designed in the framework of a rapid dialogue prototyping methodology," *Acta acustica united with acustica*, vol. 90, pp. 1906–1111, 2004.

[12] M. Turunen, J. Hakulinen, y A. Kainulainen, "Evaluation of a spoken dialogue system with usability tests and long-term pilot studies: Similarities and differences," in *Proc. of Interspeech/ICSLP 06*, Pittsburgh, USA, 2006, pp. 1057–1060.