ISSN 1870-4069

Modelo híbrido fonético-neural para corrección en sistemas de reconocimiento del habla

Rafael Viana-Cámara, Mario Campos-Soberanis,
Diego Campos-Sobrino

SoldAI Research,
México

{rviana,mcampos,dcampos}@soldai.com

Resumen. El reconocimiento automático del habla (ASR) es un área relevante en diferentes ámbitos debido a que brinda un mecanismo de comunicación natural entre aplicaciones y usuarios. A menudo los ASR presentan errores en aplicaciones que usan un léxico propio de dominios específicos. Se han explorado diversas estrategias para reducir el error en ASR cerrados mediante posprocesamiento, destacando enfoques de corrección ortográfica y aprendizaje profundo. En el presente artículo se explora el uso de una red neuronal profunda para rectificar los resultados de un algoritmo de corrección fonética, aplicado a una base de datos de audios de televenta. Los resultados obtenidos muestran una reducción de la tasa de error de palabras (WER), tanto en la transcripción original, como en la corrección fonética, mostrando la viabilidad de los modelos de aprendizaje profundo para trabajar en conjunto con estrategias de corrección posprocesamiento en la reducción de errores cometidos por ASR en dominios específicos de lenguaje.

Palabras clave: Reconocimiento del habla, corrección fonética, redes neuronales profundas.

Hybrid Phonetic-neural Model for Correction in Speech Recognition Systems

Abstract. Automatic speech recognition (ASR) is a relevant area in multiple settings because it provides a natural communication mechanism between applications and users. ASRs often fail in environments that use language specific to determined domains. Some strategies have been explored to reduce errors in closed ASRs through post-processing, in particular automatic spell checking and deep learning approaches. In this article, we explore the use of a deep neural network to refine the results of a phonetic correction algorithm, applied to a teleshopping audio database. The results exhibit a reduction in the word error rate (WER), both in the original transcription and in the phonetic correction, that shows the feasibility of deep learning models to work together with

post-processing correction strategies in the reduction of errors made by closed ASRs in specific language domains.

Keywords: Speech recognition, phonetic correction, deep neural networks.

1. Introducción

Si bien los Sistemas de Reconocimiento del Habla (ASR por sus siglas en inglés), han madurado hasta el punto de contar con algunas implementaciones comerciales de calidad, el alto rango de error que presentan en dominios específicos impiden que esta tecnología sea adoptada ampliamente[2]. Lo anterior ha motivado que la corrección en ASR haya sido abundantemente estudiada en la literatura especializada. Los ASR tradicionales se integran de tres módulos relativamente independientes: modelo acústico, modelo de diccionario y modelo de lenguaje[10]. En fechas recientes también han cobrado auge los modelos extremo a extremo de aprendizaje profundo, en los que la división modular de un sistema tradicional no es tan clara [4]. A menudo los ASR en contextos comerciales se distribuyen como cajas negras en donde los usuarios tienen poco o nulo control sobre el modelo de reconocimiento de lenguaje, lo que impide optimizarlos utilizando datos de audio propios. Esto ocasiona que los modelos de poscorrección sean el paradigma utilizado para tratar con los errores producidos por los ASR de propósito general[3]. En entornos especializados de lenguaje donde con frecuencia se encuentran términos fuera de vocabulario, el reconocimiento contextual de palabras es de suma importancia y el grado de personalización de los modelos depende de las capacidades del ASR para adaptarse al contexto. Se ha experimentado con diferentes metodologías para realizar la corrección posprocesameinto de ASR cerrados incluyendo modelos de lenguaje y corrección fonética.

En este artículo se presenta un método para la corrección posprocesamiento en sistemas ASR aplicados a dominios específicos por medio de una red neuronal de memoria a corto y largo plazo (Long Short Term Memory o LSTM) que recibe como atributos de entrada, la salida de un proceso de corrección fonética, la transcripción original del ASR y los hiperparámetros del algoritmo de corrección. A continuación se realza el aporte de la corrección neural para la generación de un algoritmo híbrido que toma en cuenta tanto la corrección fonética, como la poscorrección de la misma, lo cual resulta en una estrategia efectiva para reducir el error en el reconocimiento del habla.

El artículo está estructurado de la siguiente forma: en la sección 2 se describen antecedentes del problema y trabajos relacionados; la sección 3 presenta la metodología utilizada en la investigación; la sección 4 describe el trabajo experimental realizado presentando sus resultados en la sección 5. Finalmente se proporcionan conclusiones y líneas de experimentación para trabajo futuro en la sección 6 del artículo.

2. Antecedentes

El problema de poscorrección en ASR ha sido abordado desde diferentes perspectivas. De manera general podemos hablar de tres diferentes tipos de errores que ocurren en el reconocimiento de audio: sustitución, donde una palabra en el discurso original es transcrita como un palabra diferente; el segundo es el borrado, en el que una palabra con respecto al discurso original no se presenta en la transcripción; y finalmente, inserción en donde una palabra que no aparece en el discurso original aparece en la transcripción [2]. Existen varios esfuerzos de investigación dirigidos a realizar la corrección de los errores en ASR utilizando técnicas de posprocesamiento, en particular una cantidad importante de éstas iniciativas involucra mecanismos de retroalimentación del usuario para aprender patrones de error [2]. Entre las estrategias para aprender esos patrones de error se ha considerado la reducción del problema de poscorrección de los ASR a un problema de corrección de faltas ortográficas.

El artículo [14] propone un modelo de corrección ortográfica basado en transformadores para corregir automáticamente los errores, especialmente aquellos de sustitución realizados por un sistema de reconocimiento de voz del idioma mandarín basado en *Connectionist Temporal Classification*(CTC por sus siglas en inglés). El proyecto se llevó a cabo utilizando los resultados de reconocimiento generados por los sistemas basados en CTC como entrada y las transcripciones de verdad como salida para entrenar un transformador con arquitectura codificador-decodificador, que es muy similar a la traducción automática. Los resultados obtenidos en una tarea de reconocimiento de voz en mandarín de 20,000 horas demuestra que el modelo de corrección ortográfica propuesto en el artículo puede lograr un Character Error Rate (CER) de 3.41 %, lo que resulta en una mejora relativa de 22.9 % y 53.2 % en comparación con los sistemas de línea de base que emplean CTC decodificados con y sin modelo de lenguaje respectivamente presentados en ese mismo artículo.

Una técnica de posprocesamiento versátil basada en la distancia fonética es presentada en [11]. En dicho artículo se integra el conocimiento del dominio con los resultados de ASR de dominio abierto, lo que conduce a un mejor rendimiento. En particular, la técnica presentada es capaz de hacer uso de restricciones de dominio utilizando diversos grados de conocimiento del mismo, que van desde restricciones de vocabulario puro a través de gramáticas o n-gramas hasta restricciones de las expresiones aceptables.

Un modelo de ASR como canal de transformación ruidoso es presentado por Shivakumar et al[10] donde se propone un sistema de corrección capaz de aprender de los errores agregados de todos los módulos independientes que constituyen el ASR e intentar corregirlos. El sistema propuesto utiliza el contexto a largo plazo mediante un modelo de lenguaje de red neuronal y puede elegir de mejor manera entre las posibles transcirpciones generadas por el ASR, así como reintroducir frases previamente podadas o no vistas (que se encuentran fuera del vocabulario). Proporciona correcciones en condiciones ASR de bajo rendimiento sin degradar ninguna transcripción precisa; tales correcciones pueden incluir transcripciones fuera de dominio y no coincidentes.

El sistema expuesto en el artículo proporciona mejoras consistentes sobre el ASR de línea de base, incluso cuando éste se optimiza a través de la restauración del modelo de lenguaje de red neuronal recurrente (RNN). Los resultados demuestran que cualquier mejora de ASR puede explotarse de forma independiente y que el sistema propuesto aún puede proporcionar beneficios en sistemas de reconocimiento altamente optimizados. El beneficio del modelo de lenguaje de la red neuronal se evidencia mediante el uso de 5-gramas que permite una mejora relativa de 1.9 % sobre la línea de base-1).

En el artículo [8] se modela la distorsión en la ortografía de un nombre debido al reconocedor de voz como el efecto de un canal ruidoso. Se sigue el marco de los modelos de traducción de IBM, donde el modelo es entrenado utilizando un texto paralelo de subtítulos y salida automática de reconocimiento de voz. También se realizan pruebas con un método basado en la distancia de edición de cadena. La eficacia de los modelos se evalúa en una tarea de recuperación de consulta de nombre. Los métodos presentados en el artículo dan como resultado una mejora del 60 % en F_1.

Un modelo de incrustación de palabras robusto a ruido es propuesto en [7]. Este supera a los modelos de uso común existentes como fasttext y word2vec en diferentes tareas. Se proponen extensiones para modelos modernos en tres tareas posteriores, es decir, clasificación de texto, reconocimiento de entidades con nombre y extracción de aspectos, estas extensiones muestran una mejora en la robustez al ruido sobre las soluciones existentes para diferentes tareas de NLP.

En [1] se utilizan estrategias de corrección fonética para corregir los errores generados por un sistema ASR. El trabajo citado convierte la transcripción del ASR a una representación en formato de Alfabeto Fonético Internacional (IPA por sus siglas en inglés). Los autores emplean un algoritmo de ventana deslizante para la selección de frases candidatas para su corrección, utilizando una estrategia de selección de candidatos mediante palabras contextuales. Las palabras de dominio específico son provistas mediante un contexto generado manualmente y la distancia de edición entre su representación fonética en formato IPA. Los autores reportan una mejora en el 30 % de las frases reconocidas por el servicio ASR de Google.

En [13] se presenta una extensión del trabajo anterior, experimentando con la optimización del contexto generado mediante algoritmos genéticos. Los autores muestran el desempeño de variantes del algoritmo de corrección fonética utilizando diferentes métodos de representación y selección de candidatos, además de diferentes contextos de palabras evolucionados genéticamente a partir de las transcripciones reales de los audios. De acuerdo a los autores se observó el mejor desempeño del algoritmo de corrección fonética utilizando IPA como representación fonética y una selección incremental por letras, logrando una mejora del WER relativo de un 19 %.

El presente artículo explora un enfoque neural que rectifica las correcciones sugeridas por un algoritmo configurable de corrección fonética. Se experimentó con diversas variantes de configuración del corrector con diferentes representaciones fonéticas de las transcripciones y modificando otros parámetros.

Las correcciones propuestas por este algoritmo se evalúan mediante un clasificador generado utilizando una red neuronal LSTM con salida binaria que indica si se debe aplicar la corrección ofrecida por el algoritmo de corrección fonética. El clasificador recibe como parámetros la transcripción original del ASR, la sugerencia de corrección ofrecida por el algoritmo, así como sus hiper parámetros y calcula una salida binaria. Lo anterior se realiza con la finalidad de reducir el número de correcciones erróneas realizadas por el algoritmo, permitiendo mejorar aún más la calidad de la corrección en enfoques de ASR de caja negra sin necesidad acceder a modelos acústicos o de lenguaje generados por el ASR original.

3. Metodología

Se utilizó un algoritmo correctivo basado en la representación fonética de transcripciones generadas por el sistema de reconocimiento del habla de *Google*. Como fuente de las transcripciones se utilizaron audios recopilados de un sistema de televenta de refrescos el cual se encuentra actualmente en producción, interactuando con usuarios mexicanos. Las transcripciones reales de los ejemplos fueron utilizadas como corpus para generar ejemplos con la transcripción original del ASR, así como la corrección propuesta, etiquetados en forma binaria, donde 1 representa que la corrección propuesta debe realizarse y 0 indica lo contrario. Para el etiquetado se calculó el WER de la transcripción hipotética del ASR y el WER de la corrección propuesta. En ambos casos el WER fue calculado con respecto a la transcripción real generada por un humano y se consideró que la corrección debe realizarse cuando el WER de la versión corregida es menor al WER de la transcripción del ASR. La base de datos fue aumentada con variantes de transcripción producidas por el corrector fonético, al ser utilizado con diferentes parámetros. Esta base de datos aumentada fue utilizada para entrenar un clasificador generado mediante una red neuronal LSTM cuyo objetivo es producir una salida binaria que indica si la corrección propuesta es recomendada.

3.1. Base de datos

Los audios de ejemplo fueron recopilados durante llamadas al sistema de televentas atendido por un agente inteligente. En dichas llamadas los usuarios emitían frases requiriendo diversos productos en diferentes tamaños y presentaciones, así como expresiones naturales de una interaccción de venta (confirmación, precios, etc.). Como parte del proceso se requiere la transcripción de la voz del usuario a texto para su posterior análisis por el sistema, para ello se utiliza el servicio ASR de *Google*. La transcripción real de la frase se realizó por medio de agentes humanos y sirvió como línea de base para evaluar las transcripciones hipotéticas del ASR mediante la métrica *Word Error Rate* (WER), la cual es considerada el estándar para ASR [2].

3.2. Preprocesamiento

Con el objetivo de minimizar el efecto de diferencias lexicográficas, así como facilitar la comparación fonética entre las transcripciones hipotéticas del ASR y las transcripciones reales, fue necesario realizar un preprocesamiento de normalización de texto consistente en: limpieza de símbolos y signos de puntuación, conversión del texto a minúsculas, conversión de números a texto y expansión de abreviaturas.

La etapa de limpieza inicial tiene como objeto la eliminación de ruido existente en las transcripciones, así como la reducción de caracteres a letras y dígitos. Por su parte, las dos últimas etapas del preprocesamiento tienen el efecto de expandir el texto a una forma explícita que facilita su conversión fonética, lo cual ayuda al desempeño del corrector.

3.3. Algoritmo de corrección fonética (PhoCo)

Para el desarrollo de está investigación se utilizó el algoritmo de corrección fonética (PhoCo por sus siglas en inglés) descrito en [1,13], el cual consiste en transformar el texto transcrito a una representación fonética y comparar segmentos de ésta, con versiones fonéticas de palabras y frases comunes en el dominio de aplicación para su posible reemplazo. Estas palabras y frases son denominadas *contexto*. La comparación se realiza utilizando un umbral de similitud de distancia de Levenshtein que determina si una corrección es sugerida o no. La transcripción fonética es un sistema de símbolos gráficos que representan los sonidos del habla humana, es usado como convención para evitar las peculiaridades de cada lengua escrita y representar aquellas lenguas sin tradición escrita [6]. Entre las representaciones fonéticas utilizadas se encuentran la del Alfabeto Fonético Internacional (IPA) y una versión de worldbet (Wbet) [5] adaptada al español de México [12]. De igual manera el algoritmo permite utilizar diferentes estrategias de selección de candidatos. Para éste artículo se utilizaron las configuraciones de ventana deslizante (Win) y selección incremental por caracteres (Let) como son descritos en [13].

3.4. Clasificador neural

Para poder descubrir patrones de error en la corrección fonética, se empleó una red neuronal profunda, que recibe como entrada la transcripción original del ASR, la frase de corrección candidata proporcionada por el PhoCo junto con los hiperparámetros del algoritmo. La salida de la red neuronal es un número binario que indica si se debe realizar la corrección propuesta. Las redes neuronales, en particular las recurrentes, se han utilizado de manera efectiva en tareas de clasificación y descubrimiento de patrones de texto, por lo que se decidió modelar el proceso de rectificación para el algoritmo de corrección fonética mediante una red neuronal. La arquitectura de la red neuronal se diseñó para robustecer la detección de patrones de palabras y el seguimiento de dependencias a corto y largo plazo, por lo que se generó una topología compuesta de la siguiente forma:

– Una capa de *embeddings* de tamaño 128,
– Una capa LSTM de 60 unidades ocultas,
– Una capa de *Max pooling*,
– Una capa densa de 50 unidades ocultas,
– Una capa densa de activación sigmoide de 1 unidad.

La arquitectura utilizada se ilustra en la Fig. 1, en donde se muestra el procesamiento de las diferentes capas de la red hasta producir una salida binaria, mediante una única neurona con activación sigmoide.

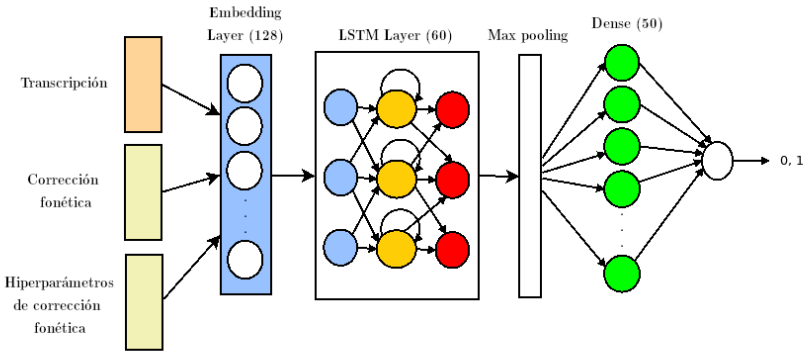


Fig. 1. Modelo del clasificador neural.

En primer lugar se tiene una capa de entrada que recibe la representación indexada del diccionario de palabras de la frase hipotética del ASR, así como de la sugerencia de corrección, y un valor numérico que indica el umbral utilizado por el PhoCo para producir su corrección candidata. Estas entradas son pasadas a una capa de *embeddings*, la cual añade una representación densa de las palabras que captura propiedades sintácticas y semánticas, las cuales han demostrado ser de utilidad en una gran cantidad de tareas de Procesamiento de Lenguaje Natural (NLP por sus siglas en inglés) [9]. A continuación se mandan las representaciones densas a una capa LSTM la cual posee propiedades importantes en el manejo de dependencias a largo plazo gracias a sus compuertas internas de actualización y olvido, que resultan de suma utilidad en la detección de patrones secuenciales de texto.

La capa de *Max pooling* funciona como un mecanismo de atención simplificado, muestreando las dependencias y entidades con mayor activación provenientes de la LSTM, propiciando la detección de características importantes en diferentes posiciones en el texto, lo cual ayuda a reducir la cantidad de datos necesarios para entrenar el modelo. A continuación se pasa por una capa densa

completamente conectada de 50 neuronas con activaciones RELU para calcular funciones compuestas por las características más relevantes muestreadas de la LSTM. Finalmente se pasa a una capa de salida de una sola neurona con función de activación sigmoide, la cual es recomendada para clasificación binaria. Se utilizó una función de pérdida de entropía cruzada binaria y se optó por una estrategia de optimización ADAM para ajustar la tasa de aprendizaje de manera adaptativa.

3.5. Algoritmo híbrido fonético-neural

El algoritmo híbrido se realizó ejecutando la corrección neuronal descrita en la sección 4.3 al algoritmo de corrección fonética, presentado en la sección 4.2. La idea central de este proceso es proveer un mecanismo de control para las posibles sustituciones erróneas que pudiera realizar el algoritmo de corrección fonética. Este enfoque permite adoptar estrategias más agresivas de corrección al poder configurar el umbral del algoritmo de corrección fonética estándar a un valor mayor y controlar los posibles errores de corrección (falsos positivos). El algoritmo consiste en realizar la corrección fonética de manera estándar y después evaluar la corrección candidata, junto con la transcripción original del ASR y los hiperparámetros del algoritmo fonético en el clasificador neural. Si el clasificador neural predice un valor mayor a 0.5 se procede a la corrección, en caso contrario se utiliza la transcripción del ASR.

4. Experimentación

En la presente sección se muestran los métodos usados para el entrenamiento del clasificador neural, la experimentación con la versión clásica del algoritmo de corrección fonética y la versión híbrida utilizando la salida del clasificador neural como factor de decisión para aceptar la corrección fonética propuesta, ilustrando los diferentes mecanismos implementados, según se describió en la sección 3 del documento.

4.1. Conjuntos de datos

Como fuente de datos para la experimentación se utilizó un total de 320 archivos de audio. Para cada uno de los audios se generaron dos transcripciones usando el ASR de Google con y sin contexto las cuales fueron almacenadas en una base de datos, conteniendo además la transcripción hecha de forma manual. Así, la base de datos contiene para cada audio dos frases hipotéticas generadas por el ASR y su transcripción real para efectos de evaluar el sistema. A continuación se realizaron diferentes hipótesis de corrección para cada ejemplo de audio utilizando diversas configuraciones del PhoCo, variando los parámetros de umbral entre 0.0 y 0.6 con un paso de 0.5, el tipo de representación utilizando IPA, texto simple y Wbet y el método de búsqueda utilizando estrategias de ventana deslizante y selección incremental de caracteres. De esta manera se

generaron 144 posibles correcciones para cada audio con lo que se consiguió una base de datos aumentada de 46,080 ejemplos para entrenar el clasificador neural. Las configuraciones listadas en la tabla se describen en [13]. Se añadió una etiqueta binaria, establecida a 1 cuando el WER de la corrección propuesta es menor al de la hipótesis del ASR y 0 en caso contrario. Los registros establecidos a 1 indican que la corrección propuesta afecta positivamente al WER.

4.2. Corrección fonética

Cada transcripción producida por el ASR en los datos de entrenamiento fue utilizada como fuente para un procedimiento de posprocesamiento correctivo basado en la transcripción fonética de texto. Dicho método de corrección fue utilizado con diferentes variantes y parámetros con lo que se obtuvieron para cada transcripción de ejemplo múltiples resultados, los cuales fueron registrados en la base de datos de entrenamiento aumentada con la estrategia presentada en la sección 4.1.

En los experimentos el parámetro de umbral fue variado usando una técnica de *GridSearch* en el rango de 0 a 0.6 en pasos de 0.05. Para el modo de representación se usaron tres variantes: IPA, texto simple y Wbet. Estas variantes en los parámetros para el corrector fonético dieron origen a variaciones en los resultados que se acumularon en la base de datos.

4.3. Clasificador neural

Para el entrenamiento del clasificador neural se dividió la base de datos aumentada descrita en la sección 4.1, en particiones aleatorias de entrenamiento, validación y prueba en porcentajes de 80 % para entrenamiento, 10 % para validación y 10 % para prueba. Se utilizó el conjunto de entrenamiento para generar diferentes modelos de redes neuronales, observando métricas de exactitud, precisión y exhaustividad, sobre los conjuntos de entrenamiento y validación, así como el área bajo la curva (AUC por sus siglas en inglés) característica operativa del receptor (ROC por sus siglas en inglés), la cual presenta un balance entre la tasa de verdaderos y falsos positivos y ofrece una métrica de desempeño para sistemas de clasificación. Se iteró sobre diferentes modelos empleando técnicas de regularización por abandono (*dropout*), con diferentes parámetros de probabilidad. Una vez se obtuvo el mejor modelo en el conjunto de validación, se procedió a evaluar el mismo en el conjunto de datos de prueba, para reportar las métricas de exactitud (*accuracy*), precisión (*precision*), exhaustividad (*recall*) y F_1 presentadas en la sección 5.1 del presente artículo. Los modelos fueron implementados utilizando tensorflow 2.0 y keras, implementados sobre un sistema opreativo Debian GNU/Linux 10 (buster) x86_64, provisto con una GPU Nvidia GTX 1080 TI de 11 GB.

4.4. Algoritmo híbrido fonético-neural

La experimentación con el algoritmo fonético neural, se realizó una vez entrenado el clasificador neural. Se procedió a examinar exhaustivamente con todos

los ejemplos de la base de datos el WER individual de cada una de las frases del ASR, del candidato a corrección fonética y de la salida del modelo fonético neural con respecto a la transcripción real. A continuación se procede a analizar el WER promedio de las frases para cada uno de los diferentes umbrales utilizados para generar la corrección fonética. En los resultados presentados en la sección 5.2 se reporta el respectivo WER promedio, así como las diferentes reducciones relativas del WER con respecto a la transcripción original.

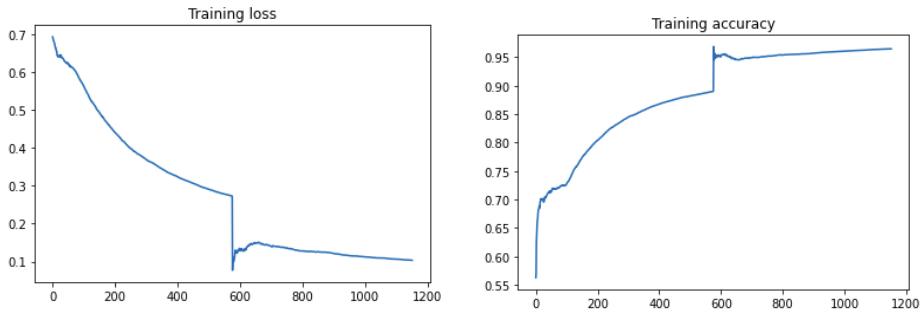
5. Resultados

En la presente sección se muestran los resultados del entrenamiento del clasificador neural, así como los comparativos entre la versión clásica del algoritmo de corrección fonética y la versión híbrida, ilustrando los diferentes valores WER promedio obtenidos de la transcripción del ASR, la corrección fonética y la corrección fonética-neural.

5.1. Clasificador neural

La red neuronal profunda se entrenó durante dos épocas con una técnica de *mini-batch* de tamaño 64, utilizando 36,863 datos obtenidos con los procedimientos descritos en la sección 4.1 y 4.3.

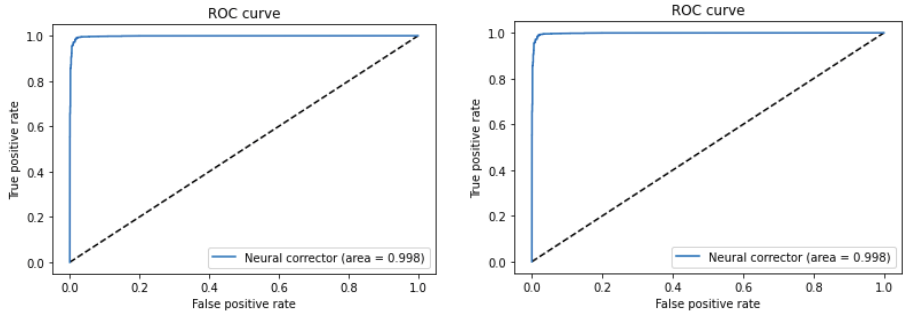
En la Fig. 2 se muestran las gráficas de la función de pérdida y la exactitud del modelo después del entrenamiento de cada uno de los lotes. La función de pérdida muestra algunas irregularidades debido a las particularidades de los diferentes lotes, sin embargo, se puede apreciar un descenso consistente en el error, en particular se nota una caída brusca alrededor del lote 550 hasta estabilizarse cerca del valor 0.1034. Un proceso similar ocurre con la exactitud de la red neuronal, la cual muestra un crecimiento sostenido, con un salto abrupto alrededor del lote 550, estabilizándose cerca del 0.9646.



(a) Pérdida (b) Exactitud

Fig. 2. Función de pérdida (a) y exactitud (b) en el entrenamiento de la red neuronal.

Una vez entrenado el mejor modelo neuronal obtenido de las diferentes fases de iteración se realizó la evaluación del mismo visualizando el área bajo la curva ROC cubierta por el modelo cuando realiza predicciones sobre los conjuntos de validación y prueba, misma que se ilustra en la Fig. 3 en donde se puede notar que se obtuvieron resultados satisfactorios cubriendo un 99 % del área.



(a) AUC validación (b) AUC prueba

Fig. 3. Área bajo la curva ROC para los conjunto de validación (a) y prueba (b).

Utilizando el modelo entrenado se calcularon métricas de exactitud, precisión, exhaustividad, así como el score F_1, utilizando el conjunto de prueba. Los resultados para las diferentes clases (0 y 1), así como el promedio realizado con la estrategia *macro average* el cual representa el promedio entre las diferentes clases. Se obtuvieron altos valores para todas las métricas superando el 95 % en cada una de ellas. El conjunto de prueba constaba de un 10 % del total de los datos traducido en 4,607 ejemplos de prueba. Los valores obtenidos para cada una de las métricas de evaluación de la red neuronal se muestra en la tabla 1, en donde llama particularmente la atención, el valor *macro average* F_1 el cual es de 98 %, siendo este un indicador de una alta eficiencia para el modelo del clasificador neural.

Tabla 1. Métricas de evaluación sobre el conjunto de datos de prueba.

Clase	**Precisión**	**Exhaustividad**	**F_1 score**	**Soporte**
0	0.99	0.99	0.99	3302
1	0.96	0.97	0.97	1305
Macro average	**0.98**	**0.98**	**0.98**	**4607**

5.2. Algoritmo híbrido fonético-neural

Los resultados de la experimentación descrita en la sección 4.3 se presentan a continuación. Los WER promedio de acuerdo a los diferentes umbrales son tomados del total de los 46,080 ejemplos. Cada uno de los valores de los umbrales utilizados para experimentar cuenta con 3,840 ejemplos. En la tabla 2, se muestran los WER promedio para los distintos umbrales así como la reducción relativa del WER para el algoritmo híbrido fonético-neural. La línea base obtenida mediante el ASR de *Google*, presentó un WER de 0.338, por lo que las reducciones relativas se realizan tomando ese valor como referencia.

Tabla 2. WER promedio y WER relativo del corrector fonético (PhoCo) y el modelo híbrido con relación al WER del ASR de Google.

Umbral	WER PhoCo	WER híbrido	$\mathbf{WER_{rel}}$ Google	$\mathbf{WER_{rel}}$ PhoCo
0.05	0.235	0.236	30.5 %	-0.1 %
0.10	0.235	0.236	30.5 %	-0.1 %
0.15	0.229	0.228	32.6 %	0.3 %
0.20	0.228	0.228	32.8 %	0.3 %
0.25	0.219	0.219	35.5 %	0.3 %
0.20	0.216	0.211	37.8 %	2.3 %
0.35	0.211	0.205	39.5 %	3.0 %
0.40	0.208	0.201	40.7 %	3.2 %
0.45	0.230	0.190	43.9 %	17.5 %
0.50	0.235	0.191	43.7 %	18.6 %
0.55	0.338	0.227	32.9 %	32.6 %
0.60	0.374	0.232	31.5 %	37.9 %
Promedio	**0.247**	**0.217**	**36.0 %**	**9.7 %**

A partir de los resultados presentados se observa que en configuraciones con umbrales pequeños (0.05 y 0.10) el WER relativo con respecto al algoritmo fonético original reduce, por lo cual el uso del clasificador neural no resulta una buena estrategia para realizar la corrección final, sin embargo, a partir de un umbral de 0.15 en adelante, muestra una mejora consistente con respecto al algoritmo fonético original, la cual aumenta notablemente conforme el valor del umbral crece, llegando a un máximo cuando el umbral también lo es y logrando una reducción del WER relativo con respecto a la versión fonética estándar de 37.9 %.

El WER relativo a la hipótesis proporcionada por el ASR de *Google*, muestra una reducción consistente, llegando a una reducción máxima de 43.9 % con un umbral del PhoCo establecido a 0.45. El algoritmo híbrido muestra reducciones consistentes del WER relativo tanto del ASR como de la transcripción fonética simple, exhibiendo una mejora promedio de 36 % y 9.7 % respectivamente. De igual manera el modelo híbrido logró obtener el WER mínimo con el umbral establecido a 0.45, disminuyendo el WER hasta 0.19, lo cual en comparación al

WER promedio del ASR de *Google*, representa una mejora del 14.8 % del WER absoluto y una del 43.9 % en términos relativos.

6. Conclusiones y trabajo futuro

A partir de los resultados obtenidos en la experimentación, se muestra la utilidad del algoritmo híbrido de corrección fonética-neural para reducir los errores de la transcripción de *Google*. Se observa que el algoritmo híbrido logra reducir el WER relativo hasta en un 43.9 %.

Se muestra una mejora consistente del algoritmo de corrección fonético-neural tanto sobre la transcripción del ASR de *Google*, como del algoritmo de corrección fonética simple. Se observó una reducción promedio del WER del algoritmo fonético simple de 9.7 %.

Las redes neuronales profundas, resultaron una excelente estrategia para el modelado de patrones de lenguaje en dominios específicos, exhibiendo un F_1 score de 0.98, así como un 99 % de área bajo la curva ROC.

Los aportes del clasificador neural resultan más notorios para valores del umbral de corrección fonética más altos, permitiendo configuraciones más agresivas para este algoritmo de corrección. Inclusive en esquemas donde el algoritmo fonético simple reduce su rendimiento debido a la corrección de ejemplos con falsos positivos, el uso a posteriori del clasificador neural resulta de utilidad para mantener un WER bajo en comparación al ASR de *Google*, como se puede observar en la tabla 2.

El corrector fonético es una estrategia viable para la corrección de errores en ASR comerciales alcanzando una mejora del WER relativo del 40.7 % con un umbral de 0.40. Con la aplicación del clasificador neural y el algoritmo híbrido se consigue reducir aún más el WER utilizando un umbral de 0.45 para el PhoCo consiguiendo una mejora en el WER relativo de 43.9 %. Este tipo de mejoras resultan importantes en ASR de uso comercial donde se necesitan grados de precisión cada vez más altos.

Dado que la arquitectura de corrección es independiente del sistema usado para la transcripción y del dominio de aplicación, la estrategia descrita es susceptible de ser extendida a diferentes sistemas ASR y dominios de aplicación. Es necesario, sin embargo, entrenar un clasificador neural para cada uno de los diferentes dominios, por lo que no se puede usar este enfoque para transferencia de conocimiento.

Los resultados encontrados muestran que es posible implementar una estrategia híbrida fonético-neural para la poscorrección de ASR en tiempo casi real. Dado que tanto el algoritmo de corrección fonética como el clasificador neural, son modelos computacionales susceptibles a escalamiento, se pueden emplear técnicas de integración de servicios web para realizar la poscorrección en sistemas ASR comerciales existentes.

Entre las líneas de investigación a futuro se requiere validar los resultados con corpus de diferentes dominios de aplicación, además se prevé la experimentación utilizando diferentes parámetros de la corrección fonética, incluyendo el

contexto y la incorporación de características del audio original. Otra línea de investigación es la comparación con algoritmos de aprendizaje profundo extremo a extremo, en donde un modelo neuronal profundo genere la corrección del ASR directamente.

Agradecimientos. A Carlos Rodrigo Castillo Sánchez, por su valiosa contribución al proporcionar la infraestructura para la experimentación del presente artículo.

Referencias

1. Campos-Sobrino, D., Campos-Soberanis, M., Martínez-Chin, I., Uc-Cetina, V.: Corrección de errores del reconocedor de voz de google usando métricas de distancia fonética. Research in Computing Science 148(1), 57–70 (2019)
2. Errattahi, R., Hannani], A.E., Ouahmane, H.: Automatic speech recognition errors detection and correction: A review. Procedia Computer Science 128, 32–37 (2018), http://www.sciencedirect.com/science/article/pii/S1877050918302187, 1st International Conference on Natural Language and Speech Processing
3. Feld, M., Momtazi, S., Freigang, F., Klakow, D., Müller, C.: Mobile texting: Can post-asr correction solve the issues? an experimental study on gain vs. costs. International Conference on Intelligent User Interfaces, Proceedings IUI (05 2012)
4. He, Y., Sainath, T.N., Prabhavalkar, R., McGraw, I., Alvarez, R., Zhao, D., Rybach, D., Kannan, A., Wu, Y., Pang, R., Liang, Q., Bhatia, D., Shangguan, Y., Li, B., Pundak, G., Sim, K.C., Bagby, T., Chang, S., Rao, K., Gruenstein, A.: Streaming end-to-end speech recognition for mobile devices. CoRR abs/1811.06621 (2018), http://arxiv.org/abs/1811.06621
5. Hieronymus, J.L.: Ascii phonetic symbols for world's languages: worldbet. Technical report, Bell Labs (1993)
6. Hualde, J.: The sounds of Spanish. Cambridge University Press (2005)
7. Malykh, V.: Robust to noise models in natural language processing tasks. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop. pp. 10–16. Association for Computational Linguistics, Florence, Italy (Jul 2019), https://www.aclweb.org/anthology/P19-2002
8. Raghavan, H., Allan, J.: Matching inconsistently spelled names in automatic speech recognizer output for information retrieval. In: Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing (01 2005)
9. Ruder, S.: A survey of cross-lingual embedding models. CoRR abs/1706.04902 (2017), http://arxiv.org/abs/1706.04902
10. Shivakumar, P.G., Li, H., Knight, K., Georgiou, P.G.: Learning from past mistakes: Improving automatic speech recognition output via noisy-clean phrase context modeling. CoRR abs/1802.02607 (2018), http://arxiv.org/abs/1802.02607
11. Twiefel, J., Baumann, T., Heinrich, S., Wermter, S.: Improving domain-independent cloud-based speech recognition with domain-dependent phonetic post-processing. In: Proceedings of the National Conference on Artificial Intelligence. vol. 2, pp. 1529–1535 (07 2014)

12. Varela, A., Cuayáhuitl, H., Nolazco-Flores, J.A.: Creating a Mexican Spanish Version of the CMU Sphinx-III Speech Recognition System. In: Sanfeliu, A., Ruiz-Shulcloper, J. (eds.) Progress in Pattern Recognition, Speech and Image Analysis. pp. 251–258. Springer Berlin Heidelberg, Berlin, Heidelberg (2003)
13. Viana-Cámara, R., Campos-Sobrino, D., Campos-Soberanis, M.: Optimización evolutiva de contextos para la corrección fonética en sistemas de reconocimiento del habla. Research in Computing Science 148(8), 293–306 (2019)
14. Zhang, S., Lei, M., Yan, Z.: Automatic spelling correction with transformer for ctc-based end-to-end speech recognition. ArXiv abs/1904.10045 (2019)