

## **REIVINDICACIONES**

1. Un método de descubrimiento automático de unigramas en un elemento de datos de voz, mediante al menos un procesador, el método comprende:

recibir un modelo lingüístico, que comprende una pluralidad de n-gramas, que comprende cada uno uno o más unigramas;

aplicar un modelo acústico de aprendizaje automático (ML) en uno o más primeros elementos de datos de voz para obtener una función de distribución de caracteres;

aplicar un decodificador voraz en la función de distribución de caracteres, para predecir un corpus inicial de unigramas;

filtrar uno o más unigramas del corpus inicial para obtener un corpus de unigramas candidatos, dichos unigramas candidatos no se incluyen en el modelo lingüístico;

analizar el uno o más primeros elementos de datos de voz, para extraer al menos un n-grama que comprende un unigrama candidato; y actualizar el modelo lingüístico para incluir al menos un n-grama extraído.

2. El método de la reivindicación 1, que comprende además aplicar un decodificador de haz en un

segundo elemento de datos de voz, para producir al menos una transcripción del segundo elemento de datos de voz, basado en el modelo lingüístico actualizado.

3. El método de la reivindicación 1, en donde  
5 la función de distribución de caracteres representa una probabilidad de aparición de uno o más caracteres lingüísticos en el uno o más primeros elementos de datos de voz.

4. El método de la reivindicación 1, que  
10 comprende además volver a entrenar el modelo acústico de ML, basado en el al menos un segundo elemento de datos de voz, usando el al menos un n-grama extraído como datos de supervisión.

5. El método de la reivindicación 1, en donde  
15 filtrar uno o más unigramas comprende además:

para uno o más unigramas candidatos,  
calcular una probabilidad de error  
ortográfico, que representa una  
probabilidad de que un unigrama relevante  
20 sea una versión con error ortográfico de un  
unigrama que ya está incluido en el modelo  
lingüístico; y  
filtrar los unigramas candidatos que  
corresponden a una probabilidad de error  
25 ortográfico que excede un umbral predefinido.

6. El método de la reivindicación 5, en donde calcular una probabilidad de error ortográfico comprende:

5            calcular un valor de distancia de Levenshtein entre el unigrama candidato y al menos un segundo unigrama, ya incluido en el modelo lingüístico;

           calcular una puntuación de frecuencia, que representa una relación de apariencia entre el unigrama candidato y el al menos un  
10            segundo unigrama en el uno o más primeros elementos de datos de voz; y

           calcular la probabilidad de error ortográfico basado en el valor de distancia de Levenshtein y la puntuación de  
15            frecuencia.

7. El método de la reivindicación 6, en donde el decodificador voraz se adapta para emitir, para cada unigrama del corpus inicial, un nivel de confianza respectivo, y donde calcular una probabilidad de error  
20            ortográfico comprende además:

           calcular una puntuación de confianza, que representa un promedio del nivel de confianza para una o más apariciones del unigrama candidato en el uno o más primeros  
25            elementos de datos de voz; y

calcular la probabilidad de error ortográfico  
basado además en la puntuación de confianza.

8. El método de la reivindicación 7, en donde  
el modelo lingüístico comprende además la definición de  
5 una o más reglas sintácticas del lenguaje, y en donde  
calcular el valor de distancia de Levenshtein comprende:

calcular un número de ediciones de un solo  
carácter entre el unigrama candidato y el  
segundo unigrama; y  
10 calcular el valor de distancia de  
Levenshtein basado en una o más reglas  
sintácticas del lenguaje y el número de  
ediciones de un solo carácter.

9. El método de la reivindicación 5, que  
15 comprende además:

para uno o más unigramas candidatos,  
calcular una probabilidad de espacio  
faltante, que representa una probabilidad  
de que el unigrama candidato sea una  
20 concatenación de dos unigramas que ya están  
incluidos en el modelo lingüístico;  
filtrar los unigramas candidatos que  
corresponden a una probabilidad de espacio  
faltante que excede un umbral predefinido.

25 10. El método de la reivindicación 7, que  
comprende además:

calcular, para al menos un primer unigrama candidato, un primer vector de incrustación de unigrama, basado en el corpus de unigramas candidatos;

5            calcular, para al menos un segundo unigrama candidato, un segundo vector de incrustación de unigrama, basado en el corpus de unigramas candidatos;

10           calcular una puntuación de similitud basada en el primer vector de incrustación de unigrama y el segundo vector de incrustación de unigrama, y

calcular la probabilidad de error ortográfico basado además en la puntuación de similitud.

15           11. El método de la reivindicación 10, que comprende además recibir un corpus de documentos, que comprende una pluralidad de documentos, en donde cada documento se asocia con un dominio de tema específico, y en donde cada documento comprende una pluralidad de  
20 unigramas de documentos.

12. El método de la reivindicación 11, que comprende además:

25           para uno o más unigramas candidatos, calcular una puntuación del dominio basada en la pluralidad de unigramas de documentos, en donde dicha puntuación del

dominio representa una probabilidad de que el unigrama candidato sea pertinente a al menos un dominio específico.

13. El método de la reivindicación 12, que  
5 comprende además, para uno o más unigramas candidatos:

compilar una lista de contexto, que comprende un subconjunto de unigramas de documentos que (a) tienen una puntuación de dominio que excede un umbral predefinido y (b) aparecen  
10 en el uno o más primeros elementos de datos de voz;

para cada documento, obtener un grupo de intersección que comprenda unigramas que aparecen en el documento y en la lista de  
15 contexto;

para cada documento, calcular una puntuación de corrección que represente la relevancia del unigrama candidato para el documento, basada en las puntuaciones del dominio de los  
20 unigramas de documentos en el grupo de intersección; y

filtrar los unigramas candidatos que corresponden a una puntuación de corrección máxima que está por debajo de un umbral  
25 predefinido.

14. El método de la reivindicación 1, en donde el decodificador voraz se adapta para emitir, para cada unigrama del corpus inicial, un nivel de confianza respectivo, y en donde analizar un elemento de datos de voz del uno o más primeros elementos de datos de voz, comprende:

para uno o más unigramas candidatos, localizar en el elemento de datos de voz, un n-grama de unigramas adyacentes, que comprende el unigrama candidato;

si (a) los unigramas de dicho n-grama corresponden a un nivel de confianza más allá de un valor predefinido, y (b) dicho n-grama comprende más que un número umbral predefinido de unigramas entonces el modelo lingüístico se actualiza para incluir el al menos un n-grama extraído; y

de cualquier otra manera, el unigrama candidato se filtra del corpus de unigramas candidatos.

15. Un sistema para descubrir automáticamente unigramas en un elemento de datos de voz, el sistema comprende: un dispositivo de memoria no transitorio, en donde se almacenan módulos de códigos de instrucciones, y al menos un procesador asociado con el dispositivo de memoria, y se configura para ejecutar los módulos de

códigos de instrucciones, con lo cual la ejecución de dichos módulos de código de instrucción, el al menos un procesador se configura para:

5            recibir un modelo lingüístico, que comprende una pluralidad de n-gramas, que comprende cada uno uno o más unigramas;

10           aplicar un modelo acústico de ML en uno o más primeros elementos de datos de voz para obtener una función de distribución de caracteres;

15           aplicar un decodificador voraz en la función de distribución de caracteres, para predecir un corpus inicial de unigramas;

20           filtrar uno o más unigramas del corpus inicial para obtener un corpus de unigramas candidatos, dichos unigramas candidatos no se incluyen en el modelo lingüístico;

             analizar el uno o más primeros elementos de datos de voz, para extraer al menos un n-grama que comprende un unigrama candidato; y

             actualizar el modelo lingüístico para incluir al menos un n-grama extraído.

16. El sistema de la reivindicación 15, que comprende además un decodificador de haz, configurado para

25   recibir al menos un segundo elemento de datos de voz, y producir una transcripción del al menos un segundo elemento

de datos de voz, basado en el modelo lingüístico actualizado.

17. El sistema de la reivindicación 15, en donde el al menos un procesador se configura para filtrar uno o  
5 más unigramas mediante:

para uno o más unigramas candidatos,  
calcular una probabilidad de error  
ortográfico, que representa una  
probabilidad de que un unigrama relevante  
10 sea una versión con error ortográfico de un  
unigrama que ya está incluido en el modelo  
lingüístico; y  
filtrar los unigramas candidatos que  
corresponden a una probabilidad de error  
15 ortográfico que excede un umbral predefinido.