(19) OFICINA ESPAÑOLA DE PATENTES Y MARCAS

ESPAÑA

(11) Número de publicación: **2 604 752**

(51) Int. CI.:

G06F 17/28 (2006.01)

(12) TRADUCCIÓN DE PATENTE EUROPEA T3

(96) Fecha de presentación y número de la solicitud europea: **20.06.2002 E 02013732 (9)**

(97) Fecha y número de publicación de la concesión europea: **07.09.2016 EP 1308851**

(54) Título: **Procedimiento de cálculo de correspondencias de traducción entre palabras de diferentes idiomas**

(30) Prioridad:

20.06.2001 US 299510 P

(45) Fecha de publicación y mención en BOPI de la traducción de la patente:
09.03.2017

(73) Titular/es:

MICROSOFT TECHNOLOGY LICENSING, LLC (100.0%)
One Microsoft Way
Redmond, WA 98052, US

(72) Inventor/es:

MOORE, ROBERT C.

(74) Agente/Representante:

CARPINTERO LÓPEZ, Mario

DESCRIPCIÓN

Procedimiento de cálculo de correspondencias de traducción entre palabras de diferentes idiomas

Antecedentes de la invención

La presente invención se refiere a aprender relaciones entre palabras. Más específicamente, la presente invención se refiere a un enfoque estadístico para aprender correspondencias de traducción entre palabras en diferentes idiomas.

Los sistemas de traducción a máquina son sistemas que reciben una entrada textual en un idioma, la traducen a un segundo idioma y proporcionan una salida textual en un segundo idioma. Al hacer esto, tales sistemas usan normalmente un lexicón de traducción para obtener correspondencias o correspondencias de traducción, entre palabras de contenido que se obtienen durante la preparación.

Un enfoque común para derivar lexicones de traducción a partir de datos empíricos implica elegir una medida de un grado de asociación entre palabras en un primer idioma, L1, y palabras en un segundo idioma, L2, en oraciones alineadas en un corpus bilingüe paralelo. Las parejas de palabras (que consisten en una palabra de L1 y una palabra de L2) se ordenan entonces por rango de acuerdo con la medida de asociación elegida. Se elige un umbral y el lexicón de traducción se forma en todas las parejas de palabras cuyo grado de asociación está por encima del umbral.

Por ejemplo, en un enfoque de la técnica anterior, la métrica de similitud (la medida del grado de asociación entre palabras) se basa en la frecuencia en que aparecen las palabras a la vez en regiones correspondientes (por ejemplo, oraciones) de un corpus de texto paralelo alineado. Las puntuaciones de selección para las diferentes parejas de palabras se computan y esas parejas de palabras se clasifican en orden descendente de su puntuación de asociación. De nuevo, se elige un umbral y las parejas de palabras cuya puntuación de asociación sobrepasa el umbral se convierten en entradas en el lexicón de la traducción.

Este tipo de procedimiento, sin embargo, tiene desventajas. El problema es que las puntuaciones de asociación se computa normalmente independientemente entre sí. Por ejemplo, se asumen que las palabras en el idioma L1 se representan por el símbolo V_k, donde k es un número entero que representa diferentes palabras en L1; y las palabras en el idioma L2 se representan por W_k, donde k es un número entero que representa diferentes palabras en L2. De esta manera, las secuencias de las V y W representan dos segmentos de texto alineados. Si W_k y V_k ocurren en contextos bilingües similares (por ejemplo, en las oraciones alineadas), entonces cualquier métrica de similitud razonable producirá una alta puntuación de asociación entre ellas, reflejando la interdependencia de sus distribuciones.

Sin embargo, se asume que V_k y V_{k+1} también aparecen en contextos similares (por ejemplo, en la misma oración). En este caso, también existe una fuerte interdependencia entre las distribuciones de V_k y V_{k+1}. De esta manera, surge el problema de que si W_k y V_k aparecen en contextos similares, y V_k y V_{k+1} aparecen en contextos similares, entonces W_k y V_{k+1} también aparecen en contextos similares. Esto se conoce como una asociación indirecta porque surge solo gracias a las asociaciones entre W_k y V_k y entre V_{k+1} y V_k. Los procedimientos anteriores que computan puntuaciones de asociación independientemente entre sí no pueden distinguir entre una asociación directa (por ejemplo, entre V_k y W_k) y una asociación indirecta (por ejemplo, entre W_k y V_{k+1}). De manera no sorprendente, esto produce lexicones de traducción repletos de asociaciones indirectas, que probablemente también son incorrectas.

Como un ejemplo concreto de una asociación indirecta, se considera el corpus paralelo de francés-inglés, que consiste principalmente en manuales de software informático traducidos. En este corpus, los términos ingleses "*file system*" y "*system files*" ocurren muy a menudo. De manera similar, los términos franceses correspondientes "*système de fichiers*" y "*fichiers système*" también aparecen juntos muy a menudo. Ya que estas colocaciones monolingües son comunes, los pares de traducción espurios *fichier/system* y *système/file* también reciben altas puntuaciones de asociación. Estas puntuaciones pueden ser mayores, de hecho, que las puntuaciones para muchos auténticos pares de traducción.

Esta deficiencia se ha abordado mediante algunas técnicas anteriores. Por ejemplo, Melamed, *Automatic Construction of Clean Broad-Coverage Translation Lexicons*, Segunda Conferencia de la Asociación para Traducción a Máquina en América (AMTA 1996), Montreal, Canadá, se dirige a este problema.

Melamed aborda este problema ignorando las parejas de palabras altamente asociadas como traducciones si se derivan de oraciones alineadas en las que existen parejas asociadas incluso más altamente que implican una o más de las mismas palabras. En otras palabras, se asume que las asociaciones más fuertes también son más fiables y de esta manera las asociaciones directas son más fuertes que las asociaciones indirectas. Por tanto, si un segmento (u oración) que contiene V se alinea con un segmento (u oración) que contiene tanto W como W', las entradas (V, W) y (V, W') no deberían aparecer en el lexicón de traducción. Si lo hacen, entonces al menos una es probablemente incorrecta. Ya que se asume que existe la tendencia de que las asociaciones directas son más fuertes que las asociaciones indirectas, entonces la entrada con la puntuación de relación más alta es la que se elige como la asociación correcta.

En el anterior ejemplo analizado, en las oraciones en inglés y francés paralelas que contienen "*fichier*" y "*système*" en el lado francés y "*file*" y "*system*" en el lado inglés, las asociaciones de *fichier/system* y *système/file* se descontarán, porque el grado de asociación para "*fichier/file*" y "*système/system*" será probablemente mucho mayor en las mismas oraciones alineadas.

Aunque este enfoque parece extender una alta precisión de salida a niveles de cobertura mucho mayores de lo que se informó anteriormente, también tiene desventajas. Por ejemplo, es muy complejo y difícil de implementar, y se cree que lleva bastante tiempo de ejecutar.

Otra dificultad encontrada en el aprendizaje de correspondencias de traducción entre las palabras implica compuestos (o secuencias de múltiples palabras que se unen para formar compuestos). Tales compuestos pueden traducirse en una única palabra en otro idioma, o múltiples palabras en el otro idioma. Las técnicas anteriores asumían que las correspondencias de traducción léxicas implicaban una única palabra. Por supuesto, tal como se muestra en la siguiente lista de compuestos, esto es manifiestamente incierto:

Base_de_donnees/database
Mot_de_passe/password
Sauvegarder/back_up
Annuler/roll_back
Ouvrir_session/log_on

En los primeros cuatro pares antes mencionados, un compuesto en un idioma se traduce como una única palabra en otro idioma. Sin embargo, en el último ejemplo, un compuesto en un idioma se traduce como un compuesto en el otro idioma, y cada uno de los componentes individuales del compuesto no puede traducirse de manera significativa en uno de los componentes individuales en el otro compuesto. Por ejemplo, "*ouvrir*", que se traduce típicamente como "*open*", no puede traducirse razonablemente ni como "*log*" ni como “*on*". De manera similar, "*session*" que se traduce normalmente como "*session*" tampoco puede traducirse razonablemente ni como "*log*" ni como "*on*".

Un intento anterior para abordar este problema también se analizó por Melamed, *Automatic Discovery of Non-Compositional Compounds in Parallel Data*, Conferencia de Procedimientos Empíricos en el Procesamiento de Idioma Natural (EMNLP 97) Providence, Rhode Island (1997). Melamed induce dos modelos de traducción, un modelo de traducción de prueba que implica un compuesto candidato y un modelo de traducción de base que no lo hace. Si el valor de la función objetiva de Melamed es mayor en el modelo de prueba que en el modelo de base, entonces el compuesto se considera válido. De lo contrario, el compuesto candidato se considera inválido. Sin embargo, el procedimiento que usa Melamed para seleccionar compuestos potenciales es bastante complejo y computacionalmente caro, ya que es su procedimiento de verificación por construcción de un modelo de traducción de prueba.

Sumario de la invención

Un corpus de preparación bilingüe paralelo se analiza sintácticamente en sus palabras de contenido. Las puntuaciones de asociación de palabras para cada pareja de palabras de contenido consisten en una palabra de un idioma L1 que ocurre en una oración alineada en el corpus bilingüe con una oración en el idioma L2 en el que ocurre la otra palabra. Una pareja de palabras se considera "vinculada" en una pareja de oraciones alineadas si una de las palabras es la que tiene la asociación más alta, de todas las palabras en esa oración, con la otra palabra. La aparición de compuestos se conjetura en los datos de preparación mediante identificación máxima, con conjuntos conectados de palabras vinculadas a cada pareja de oraciones alineadas en los datos de preparación puntuados y procesados. Siempre que uno de estos conjuntos conectados máximos contenga más de una palabra en uno o ambas idiomas, el subconjunto de las palabras en ese idioma se conjetura como un compuesto. El texto de entrada original se reescribe, sustituyendo los compuestos conjeturados por símbolos únicos fusionados. Las puntuaciones de asociación vuelven a computarse para los compuestos (que se han sustituido por símbolos fusionados) y cualquier palabra individual restante en el texto de entrada. Las puntuaciones de asociación vuelven a computarse de nuevo, excepto que esta vez, las apariciones simultáneas se tienen en cuenta al computar las puntuaciones de asociación solo donde no existe otra asociación igualmente fuerte o más fuerte en una pareja particular de oraciones alineadas en el corpus de preparación.

Las parejas de traducción pueden identificarse como aquellas parejas de palabras o parejas de símbolos que tienen puntuaciones de asociación por encima de un umbral, después de la computación final de puntuaciones de asociación.

Por supuesto, la presente invención también puede incorporarse simplemente como un procedimiento o sistema para conjeturar apariciones de compuestos en datos de preparación que comprende un corpus bilingüe alineado.

De manera similar, la descripción incluye un procedimiento de identificación de traducciones de "captoides", mediante lo que se hace referencia a títulos, u otras oraciones especiales, cuyas palabras están en mayúsculas. (Encontrar traducciones de captoides presenta un problema especial en idiomas como el francés o español, en los que la convención dicta que solo la primera palabra de tal artículo se pone en mayúsculas, por lo que la extensión de la traducción captoide es difícil de determinar). En ese ejemplo, los compuestos se identifican primero en un idioma

fuente (tal como inglés). Esto puede realizarse encontrando series de texto donde la primera palabra comienza con una mayúscula, y los últimos símbolos en la serie contigua no comienzan con una letra minúscula. A continuación, se conjeturan los compuestos en el texto meta encontrando palabras que comienzan con una mayúscula y marcando esto como el posible inicio en el compuesto correspondiente. El texto meta se escanea entonces de izquierda a derecha marcando palabras posteriores que son las que están más fuertemente relacionadas con palabras en el compuesto identificado en el texto fuente, permitiendo a la vez hasta un número predeterminado (por ejemplo, 2) de palabras contiguas no altamente relacionadas, siempre que vayan seguidas de una palabra más altamente relacionada.

El escaneo de izquierda a derecha puede continuar hasta que se encuentren más del número predeterminado (por ejemplo, más de 2) de palabras contiguas que no están más altamente relacionadas con palabras en el compuesto identificado en el texto fuente, o hasta que no haya más palabras más altamente relacionadas presentes en el texto meta, o hasta que la puntuación se alcance.

Breve descripción de los dibujos

La Figura 1 es un diagrama de bloques de un contexto general en el que la presente invención puede usarse.
La Figura 2 es un diagrama de bloques más detallado de una arquitectura de traducción a máquina general en la que la presente invención puede usarse.
La Figura 3 es un diagrama de flujo que ilustra una realización de derivar correspondencias de traducción entre palabras en un corpus bilingüe alineado.
Las Figuras. 4A-5 ilustran diferentes relaciones de asociación de palabras estadísticas entre palabras en dos idiomas diferentes.
La Figura 6 es un diagrama de flujo que ilustra una realización de identificación de compuestos conjeturados.
La Figura 7 ilustra la identificación de conjuntos conectados y máximos de palabras en parejas de oraciones alineadas en los datos de preparación.
La Figura 8 ilustra compuestos de conjetura a partir de los conjuntos conectados máximos identificados en la Figura 7.
La Figura 9 ilustra una serie de entrada reescrita usando únicos símbolos para representar compuestos conjeturados.
La Figura 10 es un diagrama de flujo que ilustra la identificación de traducciones de captoides.
La Figura 11 es un diagrama de flujo que ilustra cómo se conjeturan los compuestos correspondientes a captoides identificados.

Descripción detallada de realizaciones ilustrativas

El análisis de la Figura 1 a continuación es simplemente para exponer solo un entorno ilustrativo en el que la presente invención puede usarse, aunque puede usarse también en otros entornos.

La Figura 1 es un diagrama de bloques de un ordenador 20 de acuerdo con una realización ilustrativa de la presente invención. La Figura 1 y el análisis relacionado van destinados a proporcionar una breve descripción general de un entorno de computación adecuado en el que la invención puede implementarse. Aunque no es necesario, la invención se describirá, al menos en parte, en el contexto general de instrucciones ejecutables por ordenador, tal como módulos informáticos, que se ejecutan mediante un ordenador personal. Generalmente, los módulos de programa incluyen programas de rutina, objetos, componentes, estructuras de datos, etc., que realizan tareas particulares o implementan tipos de datos abstractos particulares. Además, los expertos en la materia apreciarán que la invención puede practicarse con otras configuraciones de sistema informático, incluyendo dispositivos manuales, sistemas de multiprocesador, sistemas electrónicos de consumidor programables o basados en microprocesador, redes PC, miniordenadores, ordenadores centrales y similares. La invención también puede practicarse en entornos de computación distribuidos donde las tareas se realizan mediante dispositivos de procesamiento remoto que se vinculan a través de una red de comunicaciones. En un entorno de computación distribuido, los módulos de programa pueden ubicarse en dispositivos de almacenamiento de memoria tanto locales como remotos.

En la Figura 1, un sistema ejemplar para implementar la invención incluye un dispositivo de computación de fin general en la forma de un ordenador 20 personal convencional, que incluye una unidad 21 de procesamiento, una memoria 22 de sistema y un bus 23 de sistema que acopla diversos componentes de sistema incluyendo la memoria del sistema con la unidad 21 de procesamiento. El bus 23 de sistema puede ser cualquiera de diversos tipos de estructura de bus que incluye un bus de memoria o un controlador de memoria, un bus periférico y un bus local usando cualquiera de una variedad de arquitecturas de bus. La memoria del sistema incluye memoria 24 de solo lectura (ROM) y memoria 25 de acceso aleatorio (RAM). Una entrada/salida 26 básica (BIOS), que contiene la rutina básica que ayuda a transferir información entre elementos dentro del ordenador 20 personal, tal como durante el inicio, se almacena en la ROM 24. El ordenador 20 personal incluye además una unidad 27 de disco duro desde la que leer y escribir en un disco duro (no se muestra), una unidad 28 de disco magnético desde la que leer o escribir en el disco 29 magnético desmontable y una unidad 30 de disco óptico desde la que leer o escribir en un disco 31 óptico desmontable tal como un CD ROM u otros medios ópticos. La unidad 27 de disco duro, la unidad 28 de disco magnético y la unidad 30 de disco óptico se conectan al bus 23 de sistema mediante una interfaz 32 de unidad de

disco duro, una interfaz 33 de unidad de disco magnético y una interfaz 34 de unidad óptica, respectivamente. Las unidades y los medios legibles por ordenador asociados proporcionan un almacenamiento no volátil de instrucciones legibles por ordenador, estructuras de datos, módulos de programa y otros datos para el ordenador 20 personal.

Aunque el entorno ejemplar descrito en el presente documento emplea un disco duro, un disco 29 magnético desmontable y un disco 31 óptico desmontable, debería apreciarse por parte de los expertos en la materia que otros tipos de medios legibles por ordenador que pueden almacenar datos que son accesibles mediante un ordenador, tal como casetes magnéticos, tarjetas de memoria flash, discos de vídeo digital, cartuchos Bernoulli, memorias de acceso aleatorio (RAM), memorias de solo lectura (ROM), y similares, también pueden usarse en el entorno operativo ejemplar.

Un número de módulos de programa pueden almacenarse en el disco duro, disco 29 magnético, disco 31 óptico, ROM 24 o RAM 25, incluyendo un sistema 35 operativo, uno o más programas 36 de aplicación, otros módulos 37 de programa y datos 38 de programa. Un usuario puede introducir órdenes e información en el ordenador 20 personal a través de dispositivos de entrada, tal como un teclado 40 y un dispositivo 42 de apuntado. Otros dispositivos de entrada (no se muestran) pueden incluir un micrófono, palanca de control, mando de juegos, antena parabólica, escáner, o similar. Estos y otros dispositivos de entrada se conectan a menudo a la unidad 21 de procesamiento a través de una interfaz 45 de puerto en serie que se acopla al bus 23 de sistema, pero que puede conectarse mediante otras interfaces, tal como una tarjeta de sonido, un puerto paralelo, un puerto de juegos o un bus en serie universal (USB). Un monitor 47 u otro tipo de dispositivo de visualización también se conecta al bus 23 de sistema mediante una interfaz, tal como un adaptador 48 de vídeo. Además del monitor 47, los ordenadores personales pueden incluir normalmente otros dispositivos de salida periféricos tales como un altavoz e impresoras (no se muestra).

El ordenador 20 personal puede funcionar en un entorno de red usando conexiones de lógica a uno o más ordenadores remotos, tales como un ordenador 49 remoto. El ordenador 49 remoto puede ser otro ordenador personal, un servidor, un router, una red PC, un dispositivo por pares u otro nódulo de red, que incluye normalmente cualquiera o todos los elementos descritos en relación con el ordenador 20 personal , aunque solo un dispositivo 50 de almacenamiento de memoria se ha ilustrado en la Figura 1. Las conexiones lógicas representadas en la Figura 1 incluyen una red 51 de área local (LAN) y una red 52 de área amplia (WAN). Tales entornos de red son lugares comunes en oficinas, intranets de red informática en todas las empresas y en Internet.

Cuando se utiliza en un entorno de red LAN, el ordenador 20 personal se conecta a la red 51 de área local a través de una interfaz o adaptador 53 de red. Cuando se usa un entorno de red WAN, el ordenador 20 personal incluye normalmente un módem 54 u otro medio para establecer comunicaciones sobre una red 52 de área amplia, tal como Internet. El módem 54, que puede ser interno o externo, se conecta al bus 23 de sistema por medio de una interfaz 46 de puerto en serie. En un entorno de red, los módulos de programa representados en relación con el ordenador 20 personal, o porciones del mismo, pueden almacenarse en los dispositivos de almacenamiento de memoria remotos. Se apreciará que las conexiones de red mostradas son ejemplares y pueden usarse otros medios de establecer un enlace de comunicaciones entre los ordenadores.

La presente invención puede utilizarse para derivar correspondencias de traducción entre palabras sustancialmente en cualquier entorno o contexto. La arquitectura de traducción a máquina que se va a describir solo es un entorno o contexto.

Aunque las formas lógicas no se necesitan para la presente invención, se analizan en relación con la arquitectura de traducción a máquina mostrada en la Figura 2. Por tanto, antes de realizar esa arquitectura en más detalle, un breve análisis de una forma lógica será útil. Un análisis completo y detallado de formas lógicas y sistemas y procedimientos para generarlas puede hallarse en la Patente de EE.UU. N.º 5.966.686 de Heidorn y col., presentada el 12 de octubre de 1999 y titulada *METHOD AND SYSTEM FOR COMPUTING SEMANTIC LOGICAL FORMS FROM SYNTAX TREES*. En resumen, sin embargo, las formas lógicas se generan realizando un análisis morfológico de una entrada de texto para producir análisis estructurales de frase convencional aumentados con relaciones gramaticales. Los análisis sintácticos sufren un procesamiento adicional para derivar formas lógicas que son estructuras de gráfico que describen dependencias etiquetadas entre palabras de contenido en la entrada textual. Las formas lógicas normalizan ciertas alternancias sintácticas, (por ejemplo, activa/pasiva) y resuelven tanto anáforas dentro de la oración como dependencias de larga distancia.

Específicamente, una relación lógica consiste en dos palabras unidas mediante un tipo de relación direccional (por ejemplo, Parte, Tiempo, Hiperónimo, Sujeto Lógico, Causa, Dominio, Ubicación, Manera, Material, Medio, Modificador, Poseedor, Fin, Cuasihiperónimo, Sinónimo, Objeto Lógico y Usuario). Una forma lógica es un gráfico de relaciones lógicas conectadas que representan una única entrada textual, tal como una oración. Esto consiste mínimamente en una relación lógica. La forma lógica representa relaciones estructurales (por ejemplo, relaciones sintácticas y semánticas), particularmente relaciones adjuntas y/o de argumentos entre palabras importantes en una serie de entrada.

En una realización ilustrativa de la arquitectura de traducción a máquina, el código particular que construye las formas lógicas a partir de análisis sintácticos se comparte por los diversos idiomas fuente y meta en los que opera el

sistema de traducción a máquina. La arquitectura compartida simplifica en gran medida la tarea de alinear los segmentos de forma lógica a partir de diferentes idiomas ya que unas construcciones superficialmente distintas en dos idiomas se colapsan frecuentemente en representaciones de forma lógica idénticas o similares.

La Figura 2 es un diagrama de bloques de una arquitectura de un sistema 200 de traducción a máquina que define una realización de un entorno para la presente invención. El sistema 200 incluye componentes 204 y 206 de análisis sintáctico, un componente 208 de aprendizaje de asociación de palabras estadístico (donde reside el volumen de la presente invención, en este entorno), un componente 210 de alineación de forma lógica, un componente 212 de construcción de base de conocimiento léxico, un diccionario 214 bilingüe, un componente 216 de fusión de diccionarios, una base 218 de datos de mapeo de transferencia y un diccionario 220 bilingüe actualizado. Durante el tiempo de ejecución, el sistema utiliza un componente 222 de análisis, un componente 224 de coincidencia, un componente 226 de transferencia y un componente 228 de generación.

En una realización ilustrativa, un corpus bilingüe se usa para preparar el sistema. El corpus bilingüe incluye oraciones traducidas alineadas (por ejemplo, oraciones en un idioma fuente o meta, tal como inglés, alineadas con sus traducciones en el otro idioma fuente o meta, tal como español o francés, etc.). Durante la preparación, las oraciones se suministran desde el corpus bilingüe alineado al sistema 200 como oraciones 230 fuente (las oraciones a traducir) y como oraciones 232 meta (la traducción de las oraciones fuente). Los componentes 204 y 206 de análisis sintáctico analizan las oraciones sintácticamente desde el corpus bilingüe alineado para producir formas 234 lógicas fuente y formas 236 lógicas meta. Durante el análisis sintáctico, las palabras en las oraciones se convierten a formas de palabras normalizadas (lemas). El término "lema" tal como se usa en el presente documento se refiere a una palabra troncal o raíz para una palabra de contenido. Por ejemplo, "dormir" es el lema para las formas superficiales "dormir", "durmiendo" y "dormido". También debería apreciarse, sin embargo, que aunque una realización de la presente invención se aplica a lemas de palabra de contenido, en otra realización, la invención puede aplicarse a formas de superficie en su lugar, pero los resultados pueden sufrir de alguna manera. En cualquier caso, los lemas se suministran después a un componente 208 de aprendizaje de asociación de palabras estadístico. Tanto las asociaciones de palabras únicas como las de múltiples palabras se conjeturan de manera iterativa y puntúan mediante el componente 208 de aprendizaje hasta que se obtiene un conjunto fiable de cada una. El componente 208 de aprendizaje de asociación de palabras estadístico envía las parejas 238 de traducción de palabras únicas aprendidas así como las parejas 240 de múltiples palabras.

Las parejas 240 de múltiples palabras se proporcionan a un componente 216 de fusión de diccionarios que se usa para añadir entradas adicionales al diccionario 214 bilingüe para formar el diccionario 220 bilingüe actualizado. Las nuevas entradas son representativas de las parejas 240 de múltiples palabras.

Las parejas 238 de palabras únicas, junto con la formas 234 lógicas fuente y las formas 236 lógicas meta, se proporcionan al componente 210 de alineación de forma lógica. El componente 210 primero establece correspondencias léxicas provisionales entre nódulos en las formas 230 y 236 lógicas fuente y meta respectivamente. Esto se realiza usando parejas de traducción a partir de un lexicón 214 bilingüe (o diccionario bilingüe) que se aumentan con las parejas 238 de traducción de palabras únicas a partir del componente 208 de aprendizaje de asociación de palabras estadístico. Después de establecer posibles correspondencias, el componente 210 de alineación alinea los nódulos de forma lógica de acuerdo tanto con características léxicas como estructurales y crea los mapeos 242 de transferencia de forma lógica.

Básicamente, el componente 210 de alineación extrae enlaces entre formas lógicas usando la información 214 de diccionario bilingüe y las parejas 238 de palabras únicas. Los mapeos de transferencia se filtran basándose en la frecuencia con la que se encuentran en las formas 234 y 236 lógicas fuente y meta y se proporcionan a un componente 212 de construcción de base de conocimiento léxico.

En un ejemplo, si el mapeo de transferencia no se ve al menos dos veces en los datos de preparación, no se usa para construir la base 218 de datos de mapeo de transferencia, aunque cualquier otra frecuencia deseada puede usarse como un filtro también. Debería apreciarse que otras técnicas de filtrado pueden usarse también, diferentes de la frecuencia de aparición. Por ejemplo, los mapeos de transferencia pueden filtrarse basándose en si se forman a partir de análisis sintácticos completos de las oraciones de entrada y basándose en si las formas lógicas usadas para crear los mapeos de transferencia se alinean completamente.

El componente 212 construye la base 218 de datos de mapeo de transferencia que contiene mapeos de transferencia que enlazan básicamente formas lógicas, o partes de las mismas, en un idioma, con formas lógicas, o partes de las mismas, en el segundo idioma. Con la base 218 de datos de mapeo de transferencia creada de esta manera, el sistema 200 se configura ahora para traducciones de tiempo de ejecución.

Durante el tiempo de ejecución, una oración 250 fuente, a traducir, se proporciona al componente 222 de análisis. El componente 222 de análisis recibe la oración 250 fuente y crea una forma 252 lógica fuente basándose en la entrada de oración fuente.

Un ejemplo puede ser útil. En el presente ejemplo, la oración 250 fuente es una oración en español "Haga clic en el botón de opción" que se traduce a inglés como "*Click the option button*" o, literalmente, "*Make click in the button of*

option".

La forma 252 lógica fuente se proporciona a un componente 224 de coincidencia. El componente 224 de coincidencia intenta hacer coincidir la forma 252 lógica fuente con las formas lógicas en la base 218 de datos de mapeo de transferencia para obtener una forma 254 lógica enlazada. Los múltiples mapeos de transferencia pueden hacer coincidir porciones de la forma 252 lógica fuente. El componente 224 de coincidencia busca el mejor conjunto de mapeos de transferencia de coincidencia en la base 218 de datos que tiene lemas de coincidencia, partes de diálogo y otra información característica. Los mapeos de transferencia más grandes (más específicos) pueden preferirse ilustrativamente a los mapeos de transferencia más pequeños (más generales). Entre los mapeos de igual tamaño, el componente 224 de coincidencia puede preferir ilustrativamente los mapeos de mayor frecuencia. Los mapeos también pueden hacer coincidir porciones de superposición de la forma 252 lógica fuente siempre que no estén en conflicto de ninguna manera.

Después de que se encuentre un conjunto óptimo de mapeo de transferencia de coincidencia, el componente 224 de coincidencia crea enlaces o nódulos en la forma 252 lógica fuente para copias de los segmentos de forma lógica meta correspondientes recibidos por los mapeos de transferencia, para generar la forma 254 lógica enlazada.

El componente 226 de transferencia recibe la forma 254 lógica enlazada desde el componente 224 de coincidencia y crea una forma 256 lógica meta que formará la base de la traducción meta. Esto se hace realizando un recorrido de arriba a abajo de la forma 254 lógica enlazada en la que se combinan los segmentos de forma lógica meta a los que apuntan los enlaces en los nódulos de la forma 252 lógica fuente. Al combinar entre sí los segmentos de forma lógica para mapeos de múltiples palabras posiblemente complejos, los subenlaces establecidos por el componente 224 de coincidencia entre nódulos individuales se usan para determinar puntos de unión correctos para modificadores, etc. Los puntos de unión por defecto se usan en caso necesario.

En casos donde no se encuentran mapeos de transferencia aplicables, los nódulos en la forma 252 lógica fuente y sus relaciones se copian simplemente en la forma 256 lógica meta. Las traducciones por defecto de palabras únicas todavía pueden encontrarse en la base 218 de datos de mapeo de transferencia para estos nódulos e insertarlas en la forma 256 lógica meta. Sin embargo, si no se encuentra ninguna, las traducciones pueden obtenerse ilustrativamente desde el diccionario 220 bilingüe actualizado que se usó durante la alineación.

El componente 228 de generación es ilustrativamente un componente de generación basado en reglas e independiente de la aplicación que mapea a partir de la forma 256 lógica meta a la serie 258 meta (u oración meta de salida). El componente 228 de generación puede no tener ilustrativamente ninguna información referente al idioma fuente de las formas lógicas de entrada, y trabaja exclusivamente con información recibida desde el componente 226 de transferencia. El componente 228 de generación también usa ilustrativamente esta información junto con un diccionario monolingüe (por ejemplo, para el idioma meta) para producir la oración 258 meta. Un componente 228 de generación genérico es de esta manera suficiente para cada idioma.

Con el anterior contexto en mente, el presente análisis continúa ahora más específicamente con respecto al componente 208 de aprendizaje de asociación de palabras estadístico. Debería apreciarse de nuevo que, aunque el presente contexto ilustra el componente 208 que funciona en formas lógicas y en una arquitectura de traducción a máquina, ese no tiene que ser necesariamente el caso. En su lugar, el componente 208 puede operar simplemente en corpus alineados que se han dividido en símbolos (o dividido en palabras individuales). El componente 208 también puede usarse para realizar otras tareas, diferentes de hacer funcionar un traductor a máquina. Por ejemplo, el componente 208 también puede usarse al formar un diccionario, o puede simplemente usarse para generar puntuaciones de asociación de palabras o relaciones entre palabras en diferentes idiomas, y no necesita funcionar en el contexto de un traductor a máquina. El anterior análisis se proporciona a modo de ejemplo únicamente.

La Figura 3 es un diagrama de flujo que ilustra un procedimiento por el que el componente 208 deriva parejas de traducción (o correspondencias de traducción entre parejas de palabras en diferentes idiomas). En primer lugar, el componente 208 obtiene acceso a un corpus bilingüe alineado. Esto se indica mediante el bloque 300. El corpus sufre un análisis sintáctico en sus palabras componentes (por ejemplo, lemas antes analizados, pero también podría mantenerse en forma de superficie). Esto se indica mediante el bloque 302. Por supuesto, en el contexto antes ilustrado, el corpus alineado sufre un análisis sintáctico mediante los componentes 204 y 206 de análisis sintáctico en formas 234 y 236 lógicas fuente y meta. Sin embargo, la presente invención no se confina a operar en entradas textuales que sufren análisis sintácticos en formas lógicas, sino que en su lugar simplemente necesita que los corpus alineados sufran un análisis sintáctico en sus palabras de contenido. Además, el analizador sintáctico también puede identificar determinados compuestos léxicos si son unidades únicas. Si tales expresiones de múltiples palabras se colocan en el lexicón, porque tienen un significado o uso específico, o porque están en uno de un número categorías generales tales como nombres propios, nombres de lugares, expresiones de tiempo, fechas, expresiones de medición, etc., se identifican como múltiples palabras.

El componente 208 a continuación computa las puntuaciones de asociación de palabras para parejas de palabras individuales en el corpus bilingüe alineado y analizado sintácticamente. Esto se indica mediante el bloque 304. Aunque puede usarse cualquier métrica de asociación de palabras que proporcione una puntuación indicativa de una asociación de palabras estadística entre parejas de palabras en el corpus de preparación, la presente invención

usa la estadística de relación de probabilidad de registro analizada por Dunning en Dunning, *Accurate Methods for the Statistics of Surprise and Coincidence, Computational Linguistics*, 19(1):61-74(1993). Esta estadística se usa para comparar la frecuencia general de una palabra o lema en el idioma 1 (WL_1) en los datos de preparación con la frecuencia de una palabra o lema en el idioma 1 (WL_1) dada una palabra o lema en el idioma 2 (WL_2) (es decir, la frecuencia con la que WL_1 ocurre en oraciones de L1 que se alinean con oraciones de L2 en las que ocurre WL_2). Al aplicar la estadística de relación de probabilidad de registro se proporciona por tanto una medición de la probabilidad de que una asociación positiva observada entre WL_1 y WL_2 no sea accidental.

La lista de parejas de palabras para las que se computan puntuaciones de asociación también puede recortarse. En otras palabras, el procedimiento de computación de las puntuaciones de asociación de palabras genera puntuaciones de asociación para un gran número de parejas de palabras (o lemas) para un corpus de preparación grande. Por tanto, en una realización ilustrativa, el conjunto de parejas de palabras se recorta para limitar adicionalmente el procesamiento a esas parejas que tienen al menos alguna posibilidad de considerarse como parejas de traducción. Una heurística ilustrativa establece este umbral como el nivel de asociación de parejas de palabras o lemas que tienen una aparición conjunta, más otra aparición cada una.

A continuación, el componente 208 conjetura la aparición de componentes en los datos de preparación y sustituye los componentes conjeturados por un único símbolo. Esto se indica mediante el bloque 306. Un ejemplo generalizado puede ser útil.

La Figura 4A muestra una secuencia de palabras en oraciones alineadas en inglés y francés. Las palabras en la secuencia en inglés se representan mediante E_x y las palabras en la secuencia en francés se representan mediante F_x. Las flechas que apuntan desde la secuencia en inglés a la secuencia en francés ilustran con qué palabras en francés están más fuertemente asociadas las palabras inglesas correspondientes. Por tanto, puede verse que E_1, por ejemplo, está más fuertemente asociada con F_1. Las flechas que apuntan desde la secuencia en francés a la secuencia en inglés ilustran cuáles de las palabras inglesas tiene una asociación más fuerte con las palabras en francés correspondientes, basándose en las puntuaciones de asociación de palabras. Por tanto, en el ejemplo, también puede verse que F_1 está más fuertemente asociada con E_1. Ya que cada una de las palabras inglesas está más fuertemente asociada con una palabra en francés correspondiente, y esa palabra en francés está más fuertemente asociada con la palabra inglesa correspondiente, se dice que existe una correspondencia simple de 1 a 1 entre la secuencia de palabras en inglés y la secuencia de palabras en francés.

De manera similar, la Figura 4B también muestra una correspondencia de 1 a 1 entre las secuencias de palabras. La Figura 4B es algo diferente a la Figura 4A porque la palabra en inglés E1 está más fuertemente asociada con la palabra en francés F2, y la palabra en inglés E2 está más fuertemente asociada con la palabra en francés F1. Sin embargo, la palabra en francés F1 también está más fuertemente asociada con la palabra en inglés E2 y la palabra en francés F2 está más fuertemente asociada con la palabra en inglés E1. Por tanto, todavía existe una correspondencia de 1 a 1 entre las secuencias de palabras, pero el orden de las palabras en francés es ligeramente diferente al orden de las palabras en inglés.

La Figura 5, sin embargo, ilustra un caso ligeramente diferente. En la Figura 5, las palabras en inglés E1 y E4 tienen una asociación de 1 a 1 con las palabras en francés F1 y F4, respectivamente. Sin embargo, aunque la palabra en inglés E2 está más fuertemente asociada con la palabra en francés F2 y la palabra en francés F2 está más fuertemente asociada con la palabra en inglés E2, la palabra en francés F3 también está más fuertemente asociada con la palabra en inglés E2. Por tanto, las palabras en inglés E2 y E3 y las palabras en francés F2 y F3 no tienen una correspondencia de 1 a 1. Esta falta de correspondencia de 1 a 1 indica en gran medida la necesidad de conjeturar compuestos para obtener traducciones correctas. Tal como se describe en más detalle con respecto a la Figura 6, las palabras en inglés E2 y E3 y las palabras en francés F2 y F3 se conjeturan como compuestos y se sustituyen por símbolos fusionados (por ejemplo, E2_E3 y F2_F3) en el texto de entrada original.

El componente 208 vuelve a computar a continuación las puntuaciones de asociación para el texto de entrada reescrito (es decir, los compuestos y cualquier palabra individual restante). Esto se indica mediante el bloque 308 en la Figura 3. Esto repite básicamente la etapa indicada en el bloque 304, con el texto reescrito en términos de los compuestos conjeturados.

A continuación, las puntuaciones de asociación se vuelven a computar de nuevo. Sin embargo, esta vez solo las apariciones conjuntas se tienen en cuenta, donde no existe otra asociación igualmente fuerte o más fuerte en las oraciones alineadas. Esto se indica en el bloque 310. En otras palabras, asumiendo que todos los compuestos necesarios para la traducción se han identificado correctamente y reformulado en los datos de preparación como un único artículo, los datos de preparación pueden tratarse como si todas las traducciones fueran de 1 a 1. Por tanto, el conjunto final de parejas de traducción clasificadas se elige asumiendo que las parejas de traducción auténticas siempre estarán mutuamente más fuertemente asociadas en una determinada pareja de oraciones alineadas. De esta manera, la recomputación de las puntuaciones de asociación indicada mediante el bloque 310 se realiza de la misma manera que la indicada en el bloque 308, excepto que las palabras en diferentes idiomas (WL_1 y WL_2) se consideran como una aparición conjunta solo si WL_1 está únicamente más fuertemente asociada con WL_2 y WL_2 está únicamente más fuertemente asociada con WL_1, entre las palabras (o lemas o lemas compuestos) presentes en una determinada pareja de oraciones alineadas. Las asociaciones computadas en la etapa 308 se usan para tomar

esta decisión. El conjunto final de asociaciones se clasifica entonces en orden decreciente de acuerdo con la fuerza de asociación.

Finalmente, aquellas parejas de palabras y/o compuestos que tienen puntuaciones de asociación por encima de un umbral en la lista final se identifican como traducciones entre sí. Esto se indica en el bloque 312. El umbral puede elegirse empíricamente, puede elegirse basándose en un análisis lingüístico de los resultados proporcionados en la lista de parejas final o puede elegirse usando otra técnica deseada.

La Figura 6 es un diagrama de flujo que ilustra, en mayor detalle, cómo los componentes se conjeturan tal como se expone en el bloque 306 en la Figura 3. Esto se analizará con respecto al ejemplo mostrado en la Figura 5 donde no existe una correspondencia directa de 1 a 1 entre las secuencias de palabras originales en las oraciones alineadas.

En primer lugar, para cada palabra en una pareja alineada de oraciones, el componente 208 identifica la palabra más fuertemente asociada en la otra oración de la pareja. En otras palabras, el componente 208 construye básicamente el gráfico o una representación del gráfico ilustrado en la Figura 5. Esto se indica mediante el bloque 320 en la Figura 6.

El componente 208 encuentra a continuación los conjuntos conectados máximos de palabras dentro de los gráficos. Esto se indica mediante el bloque 322. Básicamente, el componente 208 examina el gráfico creado para identificar áreas donde los artículos en el gráfico pueden agruparse (o rodearse) y ninguna flecha se extiende fuera de esa área. Esto se representa mediante los círculos discontinuos mostrados en la Figura 7. Cada uno de los grupos de palabras abarcado en uno de los círculos discontinuos se identifica como un conjunto conectado máximo. Todos los artículos en los conjuntos conectados máximos están de esta manera más fuertemente asociados con otro artículo en el conjunto conectado máximo, y ninguno está más fuertemente asociado con artículos fuera del conjunto conectado máximo.

El componente 208 divide entonces los conjuntos conectados máximos en los dos idiomas diferentes y conjetura que tres componentes de múltiples palabras de los conjuntos conectados máximos en cada idioma son compuestos. Esto se indica mediante el bloque 324. Por ejemplo, la Figura 8 muestra que los conjuntos conectados máximos de la Figura 7 se han dividido mediante una línea horizontal que divide la secuencia de palabras en inglés de la secuencia de palabras en francés. El componente 208 conjetura de esta manera que los componentes de múltiples palabras de los conjuntos conectados máximos en cada idioma (componentes E2 y E3 en el idioma inglés y componentes F2 y F3 en el idioma francés) son compuestos. Esto identifica compuestos, por ejemplo, tal como "*ouvir_session*" y "*log_on*".

El componente 208 reescribe entonces el archivo de entrada original sustituyendo los componentes conjeturados (E2 y E3 y F2 y F3) por símbolos fusionados. Esto se indica mediante el bloque 326. La Figura 9 ilustra esta etapa en mayor detalle. En la Figura 9, el término E2_E3 representa el símbolo en inglés correspondiente a los artículos E2 y E3 en el texto original y el artículo F2_F3 se corresponde con un símbolo que representa las palabras F2 y F3 en el texto de entrada francés original. Habiendo conjeturado y reescrito los símbolos de esta manera, el procedimiento continúa con respecto al bloque 308 en la Figura 3 donde se computan las puntuaciones de asociación de palabras para los compuestos y las palabras individuales restantes.

También puede abordarse otro problema que surge al realizar el análisis sintáctico de un texto de entrada sin procesar. En muchos tipos de texto, particularmente determinados tipos de textos técnicos, las frases no se usan de manera normal, sino que en su lugar se usan como el nombre de algo en ese dominio en particular. Por ejemplo, la oración "*Click to remove the View As Web Page check mark*" incluye el término "*View As Web Page*" que tiene la forma sintáctica de una frase en verbo no conjugado. Sin embargo, en la oración, se usa como si fuera un nombre propio. Si el analizador sintáctico no reconoce este uso especial de la frase, es virtualmente imposible analizar sintácticamente la oración correctamente.

En el idioma inglés, las expresiones de este tipo pueden manejarse de manera directa, principalmente porque las convenciones de uso de mayúsculas en inglés hacen que este tipo de frases sean fáciles de reconocer. El conversor de símbolos usado para convertir en símbolos el texto de entrada antes de analizar sintácticamente, conjetura que esas secuencias de palabras en mayúsculas, tales como "*View As Web Page*" deberían tratarse como expresiones de múltiples palabras lexicalizadas. Esta subclase de múltiples palabras se denomina en este documento "captoides".

Identificar traducciones de estos captoides, sin embargo, es muy difícil. Esto es así principalmente porque las convenciones de uso de mayúsculas en otros idiomas (tales como francés o español, por ejemplo), solo usan mayúscula en la primera palabra de tal expresión. Por tanto, aunque es relativamente directo en el idioma inglés determinar dónde comienza y termina un captoide, es muy difícil en otros idiomas.

Se proporciona un procedimiento que puede usarse para identificar traducciones de captoides y añadirlos al lexicón de traducción usado mediante el analizador sintáctico o usado en otros diversos lugares en el sistema de traducción a máquina de manera que los captoides puedan traducirse con precisión. El procedimiento se aprovecha del hecho de que, en inglés, tales captoides pueden identificarse de manera directa y también se aprovecha de las características de la presente invención que pueden usarse para identificar compuestos. La Figura 10 es un

diagrama de flujo que ilustra mejor el procedimiento de identificación de la traducción de captoides.

En primer lugar, se reciben los datos de preparación del corpus bilingüe alineado. Esto se indica mediante el bloque 350. A continuación, los datos de preparación se convierten en símbolos para obtener las diversas palabras diferentes en los datos de preparación. Cualquier conversor de símbolos disponible comercialmente puede usarse, siempre que divida los datos de preparación en palabras. Esto se indica mediante el bloque 352. A continuación, se identifican los compuestos de múltiples palabras, incluyendo captoides. Esto se indica mediante el bloque 354. En una realización, en inglés, los captoides se identifican buscando secuencias de palabras donde la primera palabra en la secuencia comienza con una mayúscula y las palabras posteriores en la secuencia no comienzan con letra minúscula. Esto permite la aparición en captoides de cosas, diferentes de letras, tales como "3,0". Una vez que los captoides se identifican, las palabras en la secuencia de palabras que conforman cada captoide se agrupan como un único símbolo por captoide. Esto se realiza colocando guiones bajos entre las palabras en cada secuencia de palabras que forma una captoide.

El componente 208 computa entonces puntuaciones de asociación de palabras o estadísticas para los símbolos enviados por el conversor 352 de símbolos, y para las palabras individuales en los captoides identificados. Las palabras individuales en cada captoide pueden identificarse de manera directa, separando simplemente los elementos del captoide en las marcas de guión bajo. La computación de las asociaciones de palabras se indica mediante el bloque 356 en la Figura 10.

El componente 208 conjetura entonces los compuestos correspondientes en el idioma meta que se corresponden con los captoides identificados en el idioma fuente mediante el conversor de símbolos. Esto se indica mediante el bloque 357. Conjeturar los compuestos que se corresponden con los captoides identificados se analizará con más detalle con respecto a la Figura 11.

El componente 208 reescribe entonces los datos de preparación sustituyendo los compuestos conjeturados por símbolos únicos. Esto se indica mediante el bloque 358 en la Figura 10.

Las puntuaciones de asociación de palabra se vuelven a computar entonces para las parejas de artículos en los datos de preparación donde cada artículo en el idioma fuente (por ejemplo, inglés) o el artículo en el idioma meta (por ejemplo, francés) es una múltiple palabra que comienza con mayúscula. Esto se indica mediante el bloque 360. Esto se debe a que el procedimiento ilustrado en la Figura 10 es para identificar traducciones de captoides. Por tanto, en la etapa 360, las puntuaciones de asociación de palabras solo deben volver a computarse para artículos donde al menos uno de los artículos en la pareja traducción es un captoide (es decir, una múltiple palabra que comienza con mayúscula). Las parejas resultantes se ordenan de acuerdo con la fuerza de su puntuación de asociación.

El componente 208 filtra entonces la lista para que incluya solo parejas de traducción donde no existe una asociación igualmente fuerte o más fuerte para cada artículo en la pareja de traducción, en todos los datos de preparación. Esto se indica mediante el bloque 362. Puede verse que las restricciones aplicadas en esta etapa son más estrictas que aquellas aplicadas, por ejemplo, en el bloque 310 de la Figura 3. Esto se debe a que, mientras que una única palabra puede tener más de una traducción en diferentes contextos, puede esperarse que la clasificación de múltiples palabras complejas representadas por un captoide reciba normalmente la misma traducción sustancialmente en todos los contextos. Por tanto, solo se aceptan las traducciones que implican captoides que están más fuertemente asociados mutuamente y únicamente por todo el corpus.

También debería apreciarse que, para centrarse en casos de mayor interés, y para incrementar la precisión, otros filtros pueden colocarse en la generación de parejas de traducción. Por ejemplo, las parejas de traducción pueden limitarse a aquellas que incluyen solo un artículo meta (tal como un artículo francés donde francés es el idioma meta) que es una de las múltiples palabras construidas en este procedimiento. De manera similar, las parejas de traducción pueden limitarse para incluir solo aquellas donde el artículo inglés es una múltiple palabra, donde todas sus palabras constituyentes están en mayúscula. Además, ya que el francés se considera generalmente como un idioma más verboso que el inglés, las parejas de traducción pueden limitarse para incluir solo aquellas donde el artículo francés contiene al menos tantas palabras como el artículo inglés. Por supuesto, estas restricciones pueden adaptarse ligeramente a otros idiomas.

De nuevo, por supuesto, al igual que con la anterior realización, puede determinarse un umbral y solamente aquellas parejas de traducción que tengan una puntuación de asociación de palabras que cumpla el umbral se consideran traducciones entre sí, y el resto pueden descartarse.

Una vez que las traducciones de los captoides se han identificado, esas traducciones se suministran ilustrativamente de vuelta a los lexicones de traducción usados por los componentes 204 y 206 de análisis sintáctico. Estas también pueden suministrarse como parejas 240 de múltiples palabras para añadirse al diccionario 204 bilingüe mediante el componente 216 de fusión de diccionarios, para obtener el diccionario 220 bilingüe actualizado.

La Figura 11 es un diagrama de flujo más detallado que ilustra cómo los componentes correspondientes a captoides identificados se conjeturan tal como se expone en el bloque 357 de la Figura 10. El procedimiento ilustrado en la Figura 11 asume que los captoides en el idioma fuente (por ejemplo, inglés) ya se han identificado. Por tanto, puede

verse que el procedimiento ilustrado en la Figura 11 es unidireccional, ya que solo intenta identificar traducciones de captoides en el idioma meta, donde los captoides ya se han identificado en el idioma fuente.

También debería apreciarse que este procedimiento de conjeturar compuestos ocurre después de que las puntuaciones de asociación de palabras se hayan computado para los símbolos que representan el texto de entrada (las palabras individuales en los captoides identificados, así como los captoides tomados como una única unidad). En una realización ilustrativa, si cualquiera de las puntuaciones de asociación entre una palabra meta (por ejemplo, una palabra en francés) y la palabra constituyente de una múltiple palabra fuente (por ejemplo, las palabras constituyentes en la múltiple palabra en inglés) son mayores que las puntuaciones de asociación entre la palabra en el idioma meta y la múltiple palabra completa en el idioma fuente, entonces la más alta de tales puntuaciones se usa para representar el grado de asociación entre la palabra en el idioma meta (por ejemplo, la palabra francesa) y la múltiple palabra en el idioma fuente (por ejemplo, la múltiple palabra en inglés).

Además, solo los conjuntos de palabras meta (por ejemplo, palabras en francés), que están más fuertemente asociados en una particular pareja de oración alineada con una múltiple palabra fuente que comienza con una palabra en mayúsculas, se reservan para su consideración como la base de los compuestos.

En este punto, el componente 208 comienza a escanear la oración en el idioma meta de la pareja alineada en consideración, de izquierda a derecha. Esto se indica en el bloque 370. El escaneo se realiza para encontrar una palabra que comience con mayúsculas. Esto se indica en el bloque 372. Si se ubica tal palabra, y es la palabra inicial en una oración, entonces se determina si es la más estrechamente relacionada con una palabra en el compuesto identificado (por ejemplo en la múltiple palabra en inglés). En ese caso, se marca como el inicio posible de un compuesto correspondiente, que es una traducción del captoide identificado. Esto se indica mediante el bloque 374 en la Figura 11.

Si la palabra ubicada en el bloque 372 es una palabra no inicial (es decir, no es la primera palabra de la oración), entonces se marca como el posible inicio de la traducción del captoide (por ejemplo, la múltiple palabra en inglés). Esto se indica en el bloque 376.

Una vez que se ubica esta primera palabra, el componente 208 continúa escaneando el texto meta de izquierda a derecha, marcando palabras posteriores que están más fuertemente relacionadas con palabras en el captoide identificado. Al hacer esto, el componente 208 permite hasta dos palabras contiguas que no están más altamente relacionadas con palabras en el captoide identificado, siempre y cuando vayan seguidas de una palabra que esté más altamente relacionada con una palabra en el captoide identificado. Esto se indica mediante el bloque 378. Esto permite que el sistema represente palabras de función (tales como palabras de función en francés) que no pueden tener altas asociaciones con nada en la múltiple palabra fuente. Siempre y cuando se cumplan estas condiciones, cada palabra posterior en la oración meta se añade a la múltiple palabra meta (la traducción del captoide identificado en el texto fuente).

El componente 208 continúa este escaneo hasta que encuentra más de dos palabras contiguas en el texto meta que no están más altamente relacionadas con palabras en el captoide identificado, o hasta que no hay más palabras en el texto meta que están más altamente relacionadas con una palabra en el captoide identificado, o hasta que se encuentra un símbolo de puntuación. Esto se indica mediante el bloque 380.

Habiendo conjeturado de esta manera los compuestos como posibles traducciones de captoides, el procedimiento continúa de nuevo en la Figura 10 en el bloque 358 donde los datos de preparación se reescriben sustituyendo los compuestos conjeturados por símbolos únicos, donde las puntuaciones de asociación se vuelven a computar y las parejas de traducción se filtran. Esto se indica en los bloques 358, 360 y 362, y se ha analizado en más detalle anteriormente.

De esta manera, puede verse que la presente invención proporciona un enfoque estadístico simplificado para derivar correspondencias de traducción entre parejas de palabras y compuestos. La presente invención ofrece ventajas sobre los sistemas anteriores ya que las realizaciones de la presente técnica son mucho menos complejas de implementar y requieren menos tiempo y recursos computacionales para ejecutarse. La presente invención también mejora la derivación de correspondencias de traducción para compuestos.

Aunque la presente invención se ha descrito en referencia a realizaciones particulares, los expertos en la materia reconocerán que pueden realizarse cambios en la forma y detalle sin apartarse del alcance de la invención.

REIVINDICACIONES

1. Un procedimiento implementado por ordenador para calcular correspondencias de traducción entre palabras, que comprende:

calcular puntuaciones de asociación de palabras para cada pareja de palabras basándose en apariciones conjuntas de palabras en cada uno de una pluralidad de conjuntos de unidades bilingües alineadas en un corpus; identificar compuestos conjeturados en las unidades basándose en las puntuaciones de asociación de palabras; y obtener correspondencias de traducción basadas en las puntuaciones de asociación de palabras recalculadas.

2. El procedimiento de la reivindicación 1 en el que las unidades bilingües y alineadas comprenden oraciones (230, 232).

3. El procedimiento de la reivindicación 1 en el que las unidades bilingües y alineadas comprenden formas (234, 236) lógicas.

4. El procedimiento de la reivindicación 1 en el que obtener correspondencias de traducción comprende:

repetir la etapa de recalcular puntuaciones de asociación de palabras considerando apariciones conjuntas de parejas, incluyendo parejas (238) de palabras, parejas (240) de compuestos y parejas de compuestos/palabras, en una pareja de unidades alineadas solo si las parejas están únicamente más fuertemente asociadas entre sí entre todas las palabras en la pareja de unidades alineadas, para obtener últimas puntuaciones de asociación de palabras.

5. El procedimiento de la reivindicación 4 en el que obtener correspondencias de traducción comprende además: clasificar parejas basándose en las últimas puntuaciones de asociación de palabras.

6. El procedimiento de la reivindicación 5 en el que obtener correspondencias de traducción comprende además: seleccionar parejas como traducciones entre sí, si las últimas puntuaciones de asociación de palabras correspondientes están por encima de un nivel de umbral.

7. El procedimiento de la reivindicación 1 en el que recalcular las puntuaciones de asociación de palabras, dados los compuestos conjeturados, comprende:

sustituir cada compuesto conjeturado por un símbolo para obtener un corpus reescrito; y
recalcular las puntuaciones de asociación de palabras en las unidades alineadas en el corpus reescrito.

8. El procedimiento de la reivindicación 1 en el que identificar compuestos conjeturados comprende:

seleccionar una pareja de unidades alineada que tiene una primera unidad en un primer idioma y una segunda unidad en un segundo idioma; e
identificar compuestos conjeturados basándose en las puntuaciones de asociación de palabras que no pueden mostrar una correspondencia de uno a uno entre palabras en la primera unidad y palabras en la segunda unidad.

9. El procedimiento de la reivindicación 8 en el que identificar compuestos conjeturados basándose en las puntuaciones de asociación de palabras que no pueden mostrar una correspondencia de uno a uno comprende:

para cada palabra en la primera unidad, identificar una palabra más fuertemente asociada en la segunda unidad; y para cada palabra en la segunda unidad, identificar una palabra más fuertemente asociada en la primera unidad.

10. El procedimiento de la reivindicación 9 en el que identificar compuestos conjeturados basándose en las puntuaciones de asociación de palabras que no pueden mostrar una correspondencia de uno a uno comprende además:

identificar conjuntos conectados máximos de palabras en las primeras y segundas unidades basándose en las palabras identificadas más fuertemente asociadas en las primeras y segundas unidades.

11. El procedimiento de la reivindicación 10 en el que identificar compuestos conjeturados comprende además:

para cada una de las primeras y segundas unidades, identificar las palabras en cada conjunto conectado máximo de múltiples palabras como un compuesto conjeturado.

12. El procedimiento de la reivindicación 1 que comprende además:

acceder al corpus antes de calcular puntuaciones de asociación de palabras.

13. El procedimiento de la reivindicación 12 que comprende además:

realizar un análisis sintáctico del corpus para obtener palabras individuales.

14. El procedimiento de la reivindicación 1 que comprende además:

después de calcular las puntuaciones de asociación de palabras, recortar parejas de palabras y que no se someten a un procesamiento adicional basándose en puntuaciones de asociación de palabras.

15. El procedimiento de la reivindicación 14 en el que recortar comprende:

retirar parejas de palabras de un procesamiento adicional si tienen una puntuación de asociación de palabras por debajo de una puntuación de umbral predeterminada.

16. El procedimiento de la reivindicación 1 en el que calcular puntuaciones de asociación de palabras comprende:

calcular las puntuaciones de asociación de palabras basándose en una forma de superficie de las palabras en cada una de las unidades bilingües alineadas.

17. El procedimiento de la reivindicación 1 en el que las palabras en cada una de las unidades bilingües alineadas se convierten en lemas antes de la etapa de calcular puntuaciones de asociación de palabras.

18. Un procedimiento implementado por ordenador de preparación de un sistema de traducción a máquina, que comprende:

obtener un corpus de unidades de múltiples palabras bilingües y alineadas;
calcular puntuaciones de asociación de palabras para parejas de palabras en el corpus basándose en la aparición conjunta de palabras en las unidades alineadas;
identificar compuestos conjeturados basándose en la ausencia de una correspondencia de uno a uno entre palabras en las unidades alineadas; y
preparar el sistema de traducción a máquina basándose en las puntuaciones de asociación de palabras y los compuestos conjeturados.

19. El procedimiento de la reivindicación 18 en el que identificar compuestos conjeturados comprende:

seleccionar una pareja de unidades alineada que tiene una primera unidad en un primer idioma y una segunda unidad en un segundo idioma; e
identificar compuestos conjeturados basándose en las puntuaciones de asociación de palabras que no pueden mostrar una correspondencia de uno a uno entre palabras en la primera unidad y palabras en la segunda unidad.

20. El procedimiento de la reivindicación 19 en el que identificar compuestos conjeturados basándose en las puntuaciones de asociación de palabras que no pueden mostrar una correspondencia de uno a uno comprende:

para cada palabra en la primera unidad, identificar una palabra más fuertemente asociada en la segunda unidad; y para cada palabra en la segunda unidad, identificar una palabra más fuertemente asociada en la primera unidad.

21. El procedimiento de la reivindicación 20 en el que identificar compuestos conjeturados basándose en las puntuaciones de asociación de palabras que no pueden mostrar una correspondencia uno a uno comprende además:

identificar los conjuntos conectados máximos de palabras en las primeras y segundas unidades basándose en las palabras identificadas más fuertemente asociadas en las primeras y segundas unidades.

22. El procedimiento de la reivindicación 21 en el que identificar compuestos conjeturados comprende además:

para cada una de las primeras y segundas unidades, identificar las palabras en cada conjunto conectado máximo de múltiples palabras como un compuesto conjeturado.

23. El procedimiento de la reivindicación 18 que comprende además, después de identificar compuestos conjeturados:

recalcular las puntuaciones de asociación de palabras, dados los compuestos conjeturados.

24. El procedimiento de la reivindicación 23 que comprende además:

repetir la etapa de recalcular las puntuaciones de asociación de palabras considerando las apariciones conjuntas de parejas, incluyendo parejas de palabras, parejas de compuestos y parejas de palabras/compuestos, en una pareja de unidades alineadas solo si las parejas están únicamente más fuertemente asociadas entre sí entre todas las palabras en la pareja de unidades alineadas, para obtener últimas puntuaciones de asociación de palabras.

25. El procedimiento de la reivindicación 24 y que comprende además:

clasificar parejas basándose en últimas puntuaciones de asociación de palabras.

26. El procedimiento de la reivindicación 25 que comprende además:

seleccionar parejas como traducciones entre sí, si las últimas puntuaciones de asociación de palabras correspondientes están por encima de un nivel de umbral.

27. El procedimiento de la reivindicación 23 en el que recalcular las puntuaciones de asociación de palabras, dados los compuestos conjeturados, comprende:

sustituir cada compuesto conjeturado por un símbolo para obtener un corpus reescrito; y
recalcular las puntuaciones de asociación de palabras en las unidades alineadas en el corpus reescrito.

28. El procedimiento de la reivindicación 26 en el que la preparación del sistema de traducción a máquina, basado en las puntuaciones de asociación de palabras y los compuestos conjeturados, comprende:

generar mapeos de transferencia que mapean una unidad en uno de los idiomas a una unidad en el otro de los idiomas basándose en las traducciones seleccionadas.

29. El procedimiento de la reivindicación 18 que comprende además:

convertir las palabras a lemas antes de calcular las puntuaciones de asociación de palabras.

30. El procedimiento de la reivindicación 18 en el que las palabras son formas de superficie de las palabras.

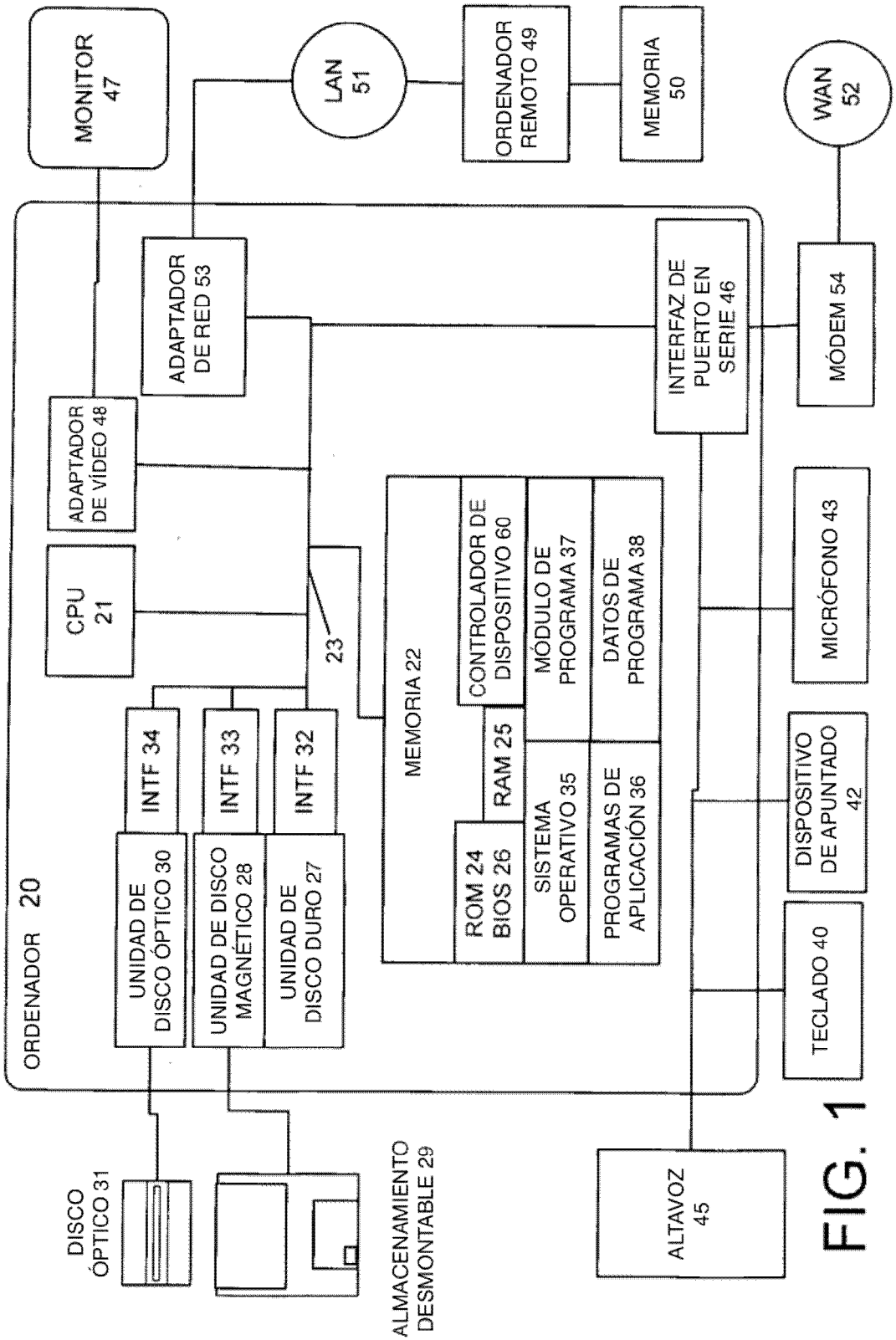
MONITOR 47
LAN 51
ORDENADOR REMOTO 49
MEMORIA 50
WAN 52
ADAPTADOR DE RED 53
INTERFAZ DE PUERTO EN SERIE 46
MÓDEM 54
ADAPTADOR DE VÍDEO 48
CPU 21
23
MEMORIA 22
CONTROLADOR DE DISPOSITIVO 60
MÓDULO DE PROGRAMA 37
DATOS DE PROGRAMA 38
MICRÓFONO 43
INTF 34
INTF 33
INTF 32
RAM 25
ROM 24 BIOS 26
SISTEMA OPERATIVO 35
PROGRAMAS DE APLICACIÓN 36
DISPOSITIVO DE APUNTADO 42
ORDENADOR 20
UNIDAD DE DISCO ÓPTICO 30
UNIDAD DE DISCO MAGNÉTICO 28
UNIDAD DE DISCO DURO 27
TECLADO 40
DISCO ÓPTICO 31
ALMACENAMIENTO DESMONTABLE 29
ALTAVOZ 45

FIG. 1

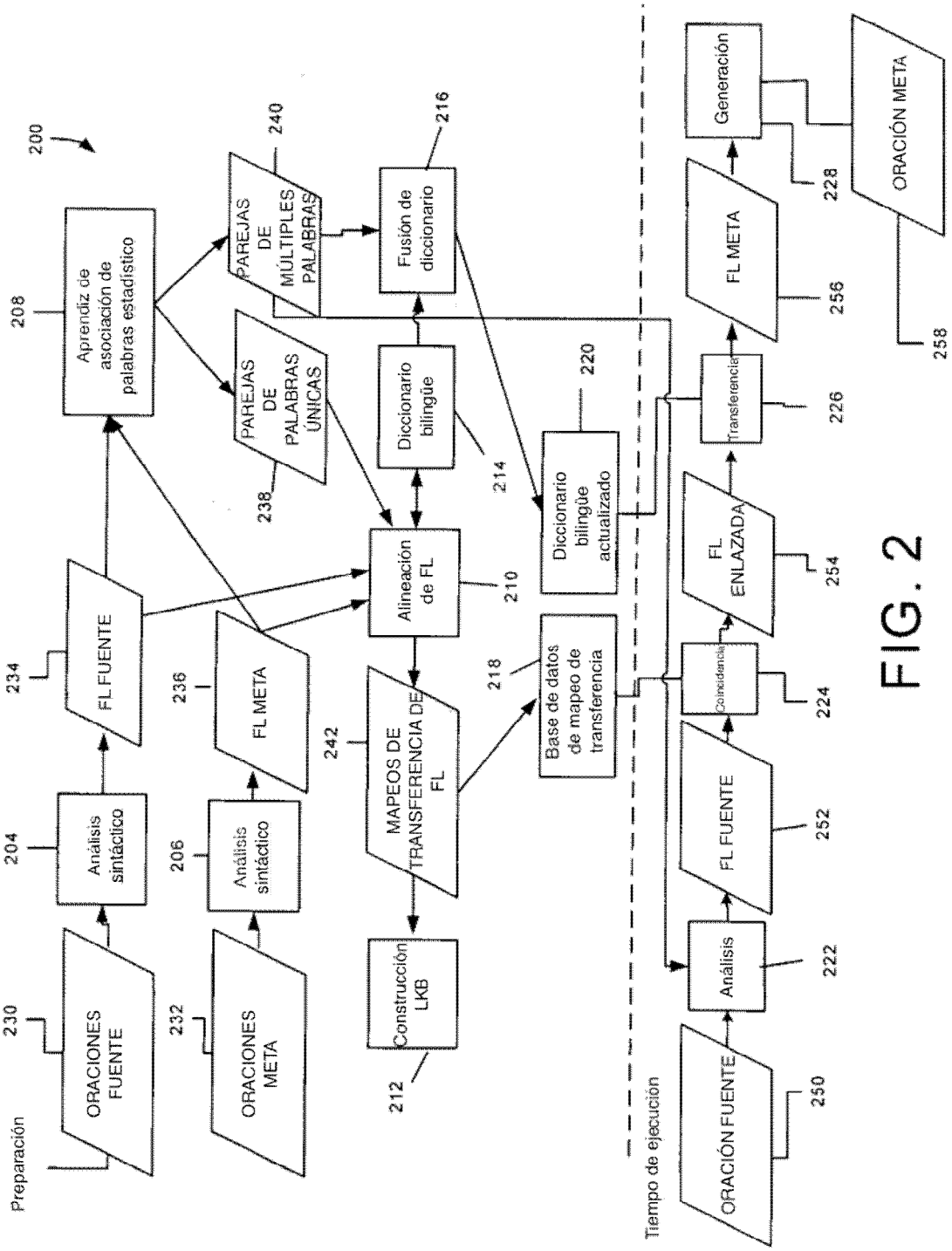
200
Preparación
Tiempo de ejecución
ORACIONES FUENTE
230
ORACIONES META
232
Análisis sintáctico
204
Análisis sintáctico
206
FL FUENTE
234
FL META
236
Aprendiz de asociación de palabras estadístico
208
PAREJAS DE PALABRAS ÚNICAS
238
PAREJAS DE MÚLTIPLES PALABRAS
240
Fusión de diccionario
216
Diccionario bilingüe
214
Alineación de FL
210
MAPEOS DE TRANSFERENCIA DE FL
242
Construcción LKB
212
Diccionario bilingüe actualizado
220
Base de datos de mapeo de transferencia
218
ORACIÓN FUENTE
250
Análisis
222
FL FUENTE
252
Coincidencia
224
FL ENLAZADA
254
Transferencia
226
FL META
256
Generación
228
ORACIÓN META
258

FIG. 2

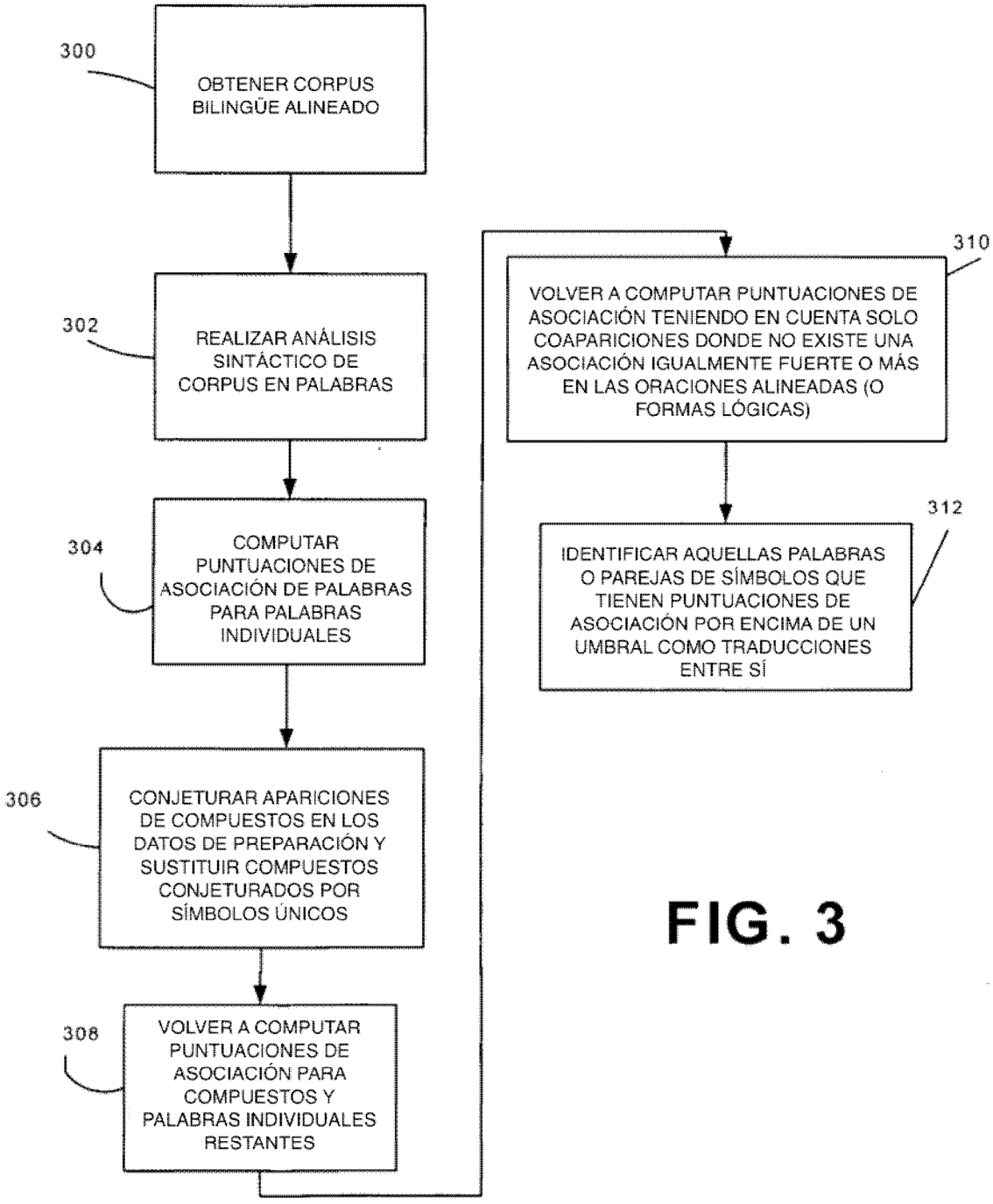
300
OBTENER CORPUS BILINGÜE ALINEADO
302
REALIZAR ANÁLISIS SINTÁCTICO DE CORPUS EN PALABRAS
304
COMPUTAR PUNTUACIONES DE ASOCIACIÓN DE PALABRAS PARA PALABRAS INDIVIDUALES
306
CONJETURAR APARICIONES DE COMPUESTOS EN LOS DATOS DE PREPARACIÓN Y SUSTITUIR COMPUESTOS CONJETURADOS POR SÍMBOLOS ÚNICOS
308
VOLVER A COMPUTAR PUNTUACIONES DE ASOCIACIÓN PARA COMPUESTOS Y PALABRAS INDIVIDUALES RESTANTES
310
VOLVER A COMPUTAR PUNTUACIONES DE ASOCIACIÓN TENIENDO EN CUENTA SOLO COAPARICIONES DONDE NO EXISTE UNA ASOCIACIÓN IGUALMENTE FUERTE O MÁS EN LAS ORACIONES ALINEADAS (O FORMAS LÓGICAS)
312
IDENTIFICAR AQUELLAS PALABRAS O PAREJAS DE SÍMBOLOS QUE TIENEN PUNTUACIONES DE ASOCIACIÓN POR ENCIMA DE UN UMBRAL COMO TRADUCCIONES ENTRE SÍ

FIG. 3

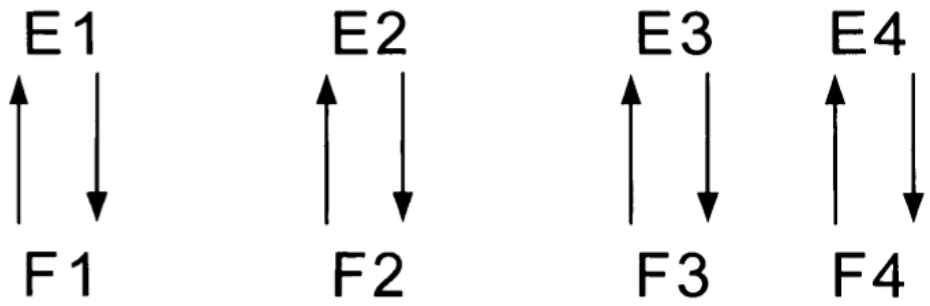

FIG. 4A

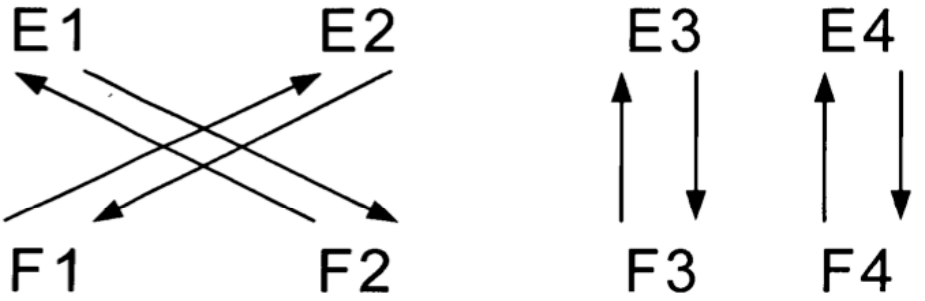

FIG. 4B

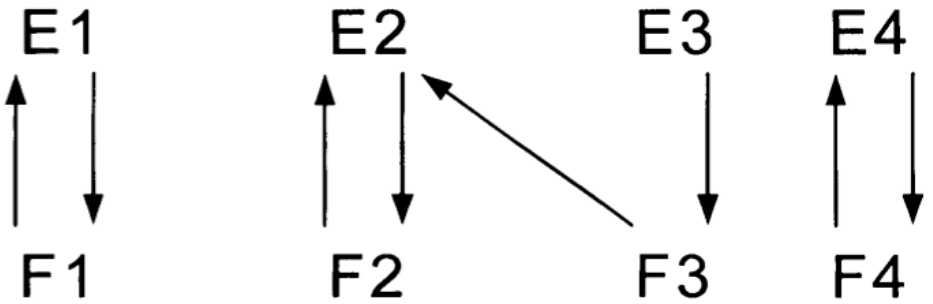

FIG. 5

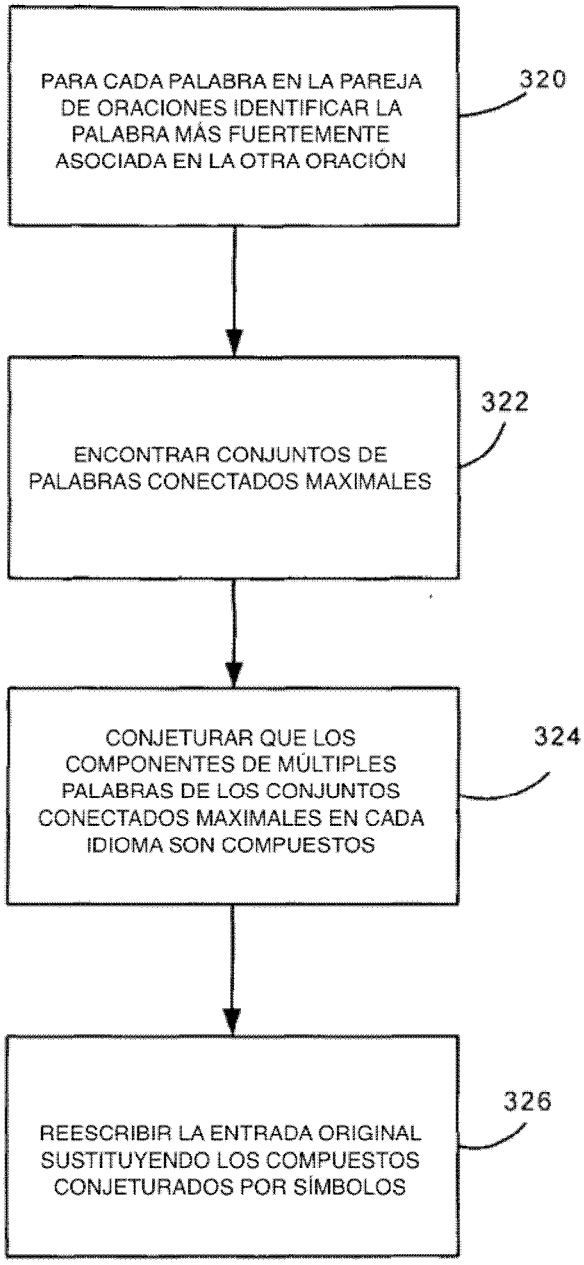

FIG. 6

E1 E2__E3 E4
F1 F2__F3 F4

FIG. 9

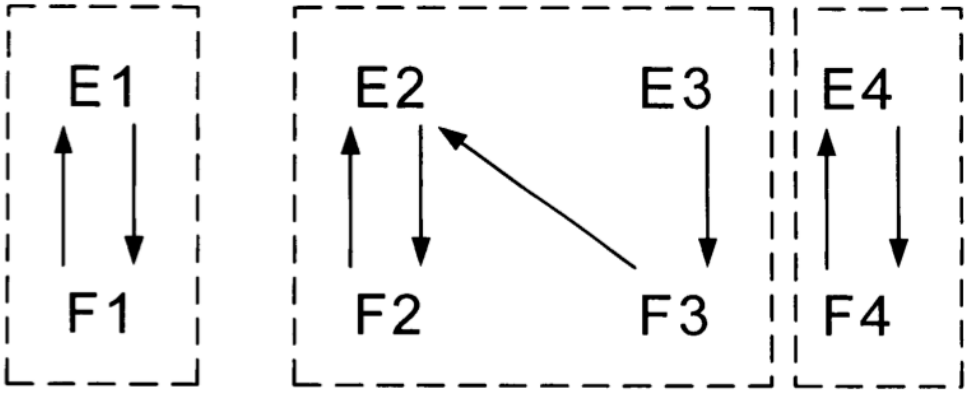

FIG. 7

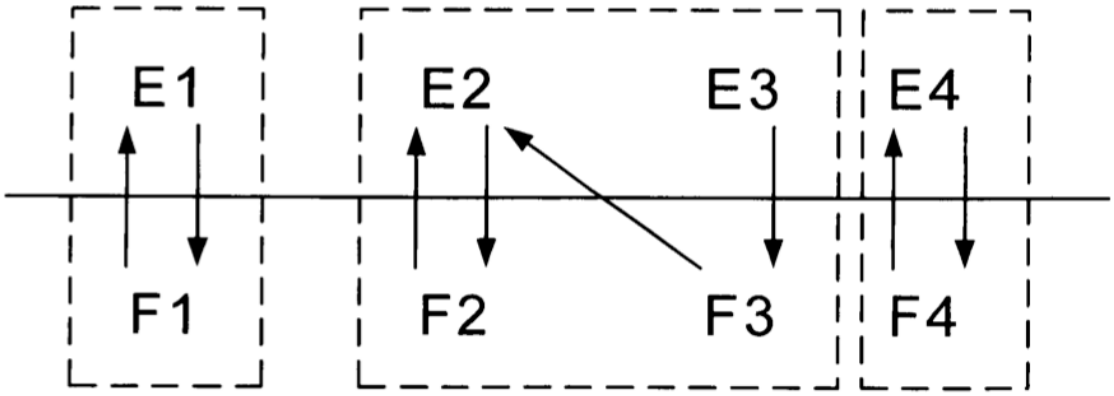

FIG. 8

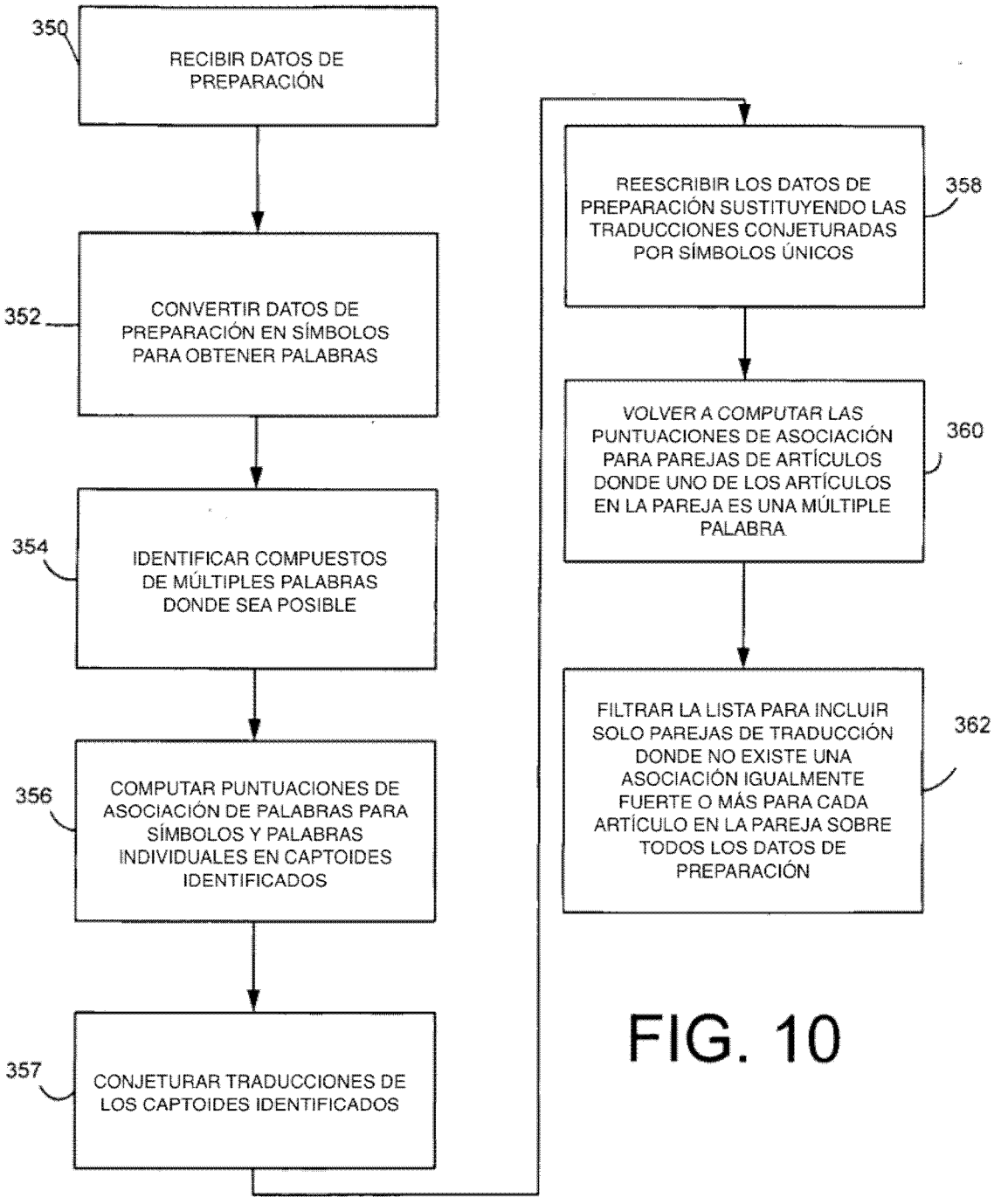
350
RECIBIR DATOS DE PREPARACIÓN
352
CONVERTIR DATOS DE PREPARACIÓN EN SÍMBOLOS PARA OBTENER PALABRAS
354
IDENTIFICAR COMPUESTOS DE MÚLTIPLES PALABRAS DONDE SEA POSIBLE
356
COMPUTAR PUNTUACIONES DE ASOCIACIÓN DE PALABRAS PARA SÍMBOLOS Y PALABRAS INDIVIDUALES EN CAPTOIDES IDENTIFICADOS
357
CONJETURAR TRADUCCIONES DE LOS CAPTOIDES IDENTIFICADOS
358
REESCRIBIR LOS DATOS DE PREPARACIÓN SUSTITUYENDO LAS TRADUCCIONES CONJETURADAS POR SÍMBOLOS ÚNICOS
360
VOLVER A COMPUTAR LAS PUNTUACIONES DE ASOCIACIÓN PARA PAREJAS DE ARTÍCULOS DONDE UNO DE LOS ARTÍCULOS EN LA PAREJA ES UNA MÚLTIPLE PALABRA
362
FILTRAR LA LISTA PARA INCLUIR SOLO PAREJAS DE TRADUCCIÓN DONDE NO EXISTE UNA ASOCIACIÓN IGUALMENTE FUERTE O MÁS PARA CADA ARTÍCULO EN LA PAREJA SOBRE TODOS LOS DATOS DE PREPARACIÓN

FIG. 10

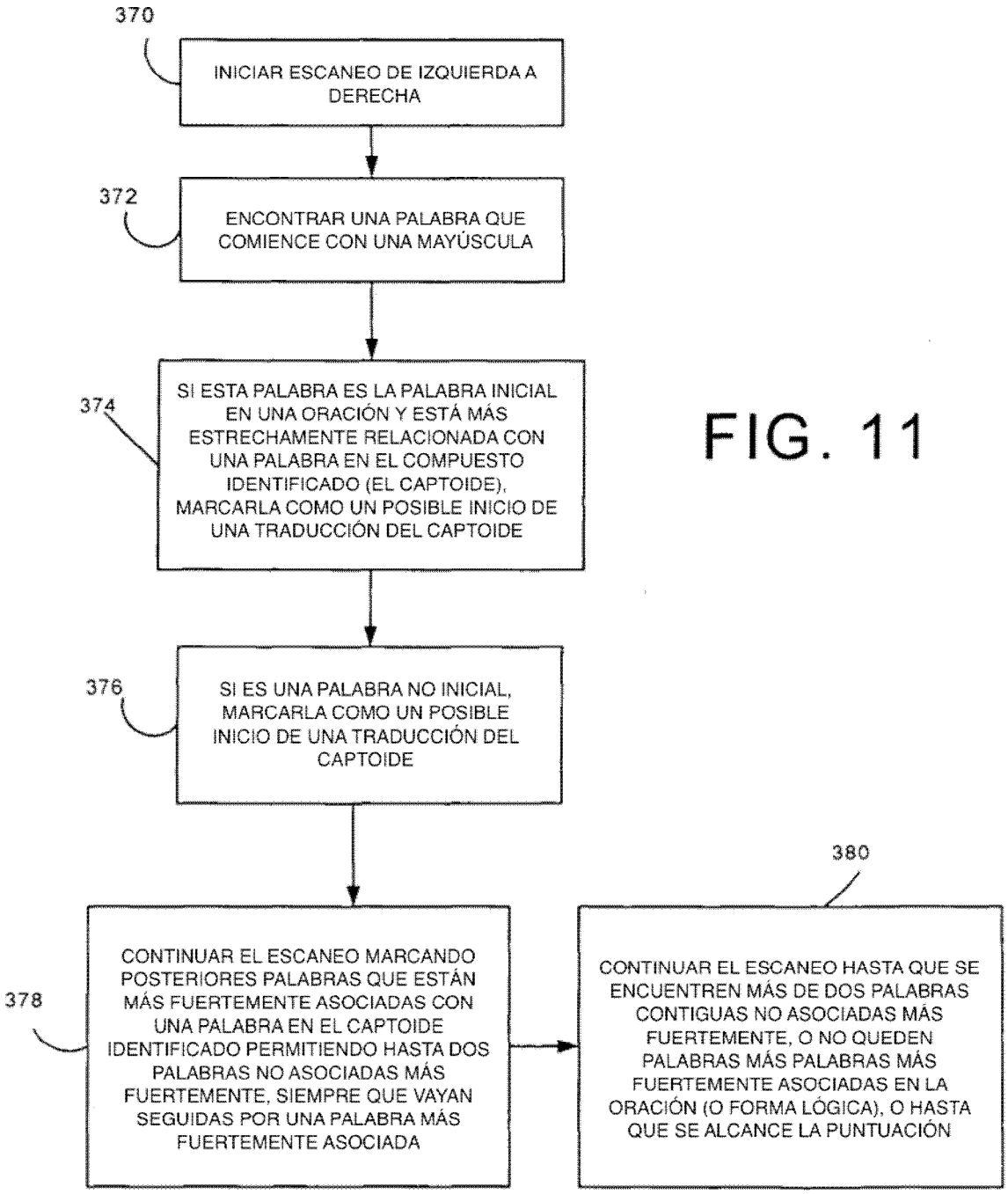
370
INICIAR ESCANEO DE IZQUIERDA A DERECHA
372
ENCONTRAR UNA PALABRA QUE COMIENCE CON UNA MAYÚSCULA
374
SI ESTA PALABRA ES LA PALABRA INICIAL EN UNA ORACIÓN Y ESTÁ MÁS ESTRECHAMENTE RELACIONADA CON UNA PALABRA EN EL COMPUESTO IDENTIFICADO (EL CAPTOIDE), MARCARLA COMO UN POSIBLE INICIO DE UNA TRADUCCIÓN DEL CAPTOIDE
376
SI ES UNA PALABRA NO INICIAL, MARCARLA COMO UN POSIBLE INICIO DE UNA TRADUCCIÓN DEL CAPTOIDE
378
CONTINUAR EL ESCANEO MARCANDO POSTERIORES PALABRAS QUE ESTÁN MÁS FUERTEMENTE ASOCIADAS CON UNA PALABRA EN EL CAPTOIDE IDENTIFICADO PERMITIENDO HASTA DOS PALABRAS NO ASOCIADAS MÁS FUERTEMENTE, SIEMPRE QUE VAYAN SEGUIDAS POR UNA PALABRA MÁS FUERTEMENTE ASOCIADA
380
CONTINUAR EL ESCANEO HASTA QUE SE ENCUENTREN MÁS DE DOS PALABRAS CONTIGUAS NO ASOCIADAS MÁS FUERTEMENTE, O NO QUEDEN PALABRAS MÁS PALABRAS MÁS FUERTEMENTE ASOCIADAS EN LA ORACIÓN (O FORMA LÓGICA), O HASTA QUE SE ALCANCE LA PUNTUACIÓN

FIG. 11