(19) OFICINA ESPAÑOLA DE PATENTES Y MARCAS

ESPAÑA

(11) Número de publicación: **2 416 359**

(51) Int. CI.:

G06F 17/28 (2006.01)

(12) TRADUCCIÓN DE PATENTE EUROPEA T3

(96) Fecha de presentación y número de la solicitud europea: **23.03.2005 E 05102340 (6)**

(97) Fecha y número de publicación de la concesión europea: **08.05.2013 EP 1582997**

(54) Título: **Traducción automática usando formas lógicas**

(30) Prioridad:

30.03.2004 US 813208

(45) Fecha de publicación y mención en BOPI de la traducción de la patente:
31.07.2013

(73) Titular/es:

MICROSOFT CORPORATION (100.0%)
ONE MICROSOFT WAY
REDMOND, WASHINGTON 98052-6399, US

(72) Inventor/es:

AUE, ANTHONY;
MENEZES, ARUL A.;
QUIRK, CHRISTOPHER B.;
RINGGER, ERIC K. y
MOORE, ROBERT C.

(74) Agente/Representante:

CARPINTERO LÓPEZ, Mario

DESCRIPCIÓN

Traducción automática usando formas lógicas

Antecedentes de la invención

La presente invención se refiere a sistemas de lenguaje automatizados. Más específicamente, la presente invención se refiere a modelos de lenguaje en sistemas de lenguaje estadísticos.

Los sistemas de lenguaje automatizados incluyen el reconocimiento del habla, el reconocimiento de la escritura manual, la producción del habla, la corrección gramatical y la traducción automática.

Los sistemas de traducción automática (MT) son sistemas que reciben una entrada en un idioma (un idioma "de origen"), traducen la entrada a un segundo idioma (un idioma "de destino") y proporcionan una salida en el segundo idioma.

Un ejemplo de un sistema de MT usa formas lógicas (LF), que son gráficos de dependencia que describen dependencias etiquetadas entre palabras del contenido en una cadena, como un paso intermedio en la traducción. Según este sistema, una cadena en el idioma de origen es analizada primero con un analizador sintáctico del idioma natural para producir una LF de origen. La LF de origen debe luego ser convertida en una LF del idioma de destino. Una base de datos de correlaciones entre trozos de LF del idioma de origen y trozos de LF del idioma de destino (junto con otros metadatos, tales como los tamaños de las correlaciones y las frecuencias de las correlaciones en algunos conjuntos de entrenamiento) es usada para esta conversión. Habitualmente, el trozo de LF del idioma de origen de una única correlación no cubre la LF de origen entera. Como resultado, debe seleccionarse un conjunto de correlaciones (posiblemente solapadas) y sus trozos de LF del idioma de destino deben ser combinados para formar una LF de destino completa.

Para identificar el conjunto de formas lógicas de destino, un sistema de MT usa un algoritmo de búsqueda voraz para seleccionar una combinación de correlaciones a partir de las posibles correlaciones cuyos trozos de LF del idioma de origen coincidan con la LF de origen. Esta búsqueda voraz comienza clasificando las correlaciones por tamaño, frecuencia y otras características que miden cuán bien coinciden los trozos de LF del idioma de origen de la correlación con la LF de origen. La lista clasificada es luego recorrida de arriba hacia abajo y se escoge el primer conjunto de correlaciones compatibles hallado que cubra la forma lógica de origen. Este sistema heurístico, sin embargo, no prueba todas las posibles combinaciones de correlaciones de entrada, sino que sencillamente selecciona el primer conjunto de correlaciones que cubra completamente la LF de origen.

Después de que está seleccionado el conjunto de correlaciones, los trozos de LF del idioma de destino de las correlaciones son combinados de una manera congruente con la LF de origen, para producir una LF de destino. Finalmente, la ejecución de un sistema de generación de idioma natural en la LF de destino produce la salida en el idioma de destino.

Sin embargo, los sistemas de MT no siempre emplean formas lógicas u otras estructuras sintácticamente analizadas como representaciones intermedias. Ni tampoco usan necesariamente procedimientos heurísticos para resolver ambigüedades de traducción. Algunos otros sistemas de MT intentan predecir la cadena del idioma de salida más probable, dada una cadena de entrada en el idioma de origen, usando modelos estadísticos. Tales sistemas de MT usan marcos y modelos estadísticos tradicionales, tales como el marco de canal ruidoso, para descodificar y hallar la oración *T* de destino que sea la más probable traducción para una oración S dada de origen. La maximización de esta probabilidad está representada por la:

Ecuación 1

$$T = \arg\max_{T'} P(T' \mid S)$$

donde *T'* varía sobre las oraciones en el idioma de destino. Usando la Regla de Bayes, la maximización de esta probabilidad también puede ser representada por:

Ecuación 2

$$T = \arg\max_{T'} P(S \mid T') \times P(T')$$

donde $P(S \mid T')$ es la probabilidad de la cadena S de origen, dada una cadena *T'* del idioma de destino, y $P(T')$ es la probabilidad de la cadena *T'* del idioma de destino. En la MT estadística basada en cadenas (MT donde no se usa

ninguna representación intermedia sintácticamente analizada), se usa un modelo de idioma de destino entrenado en datos monolingüísticos del idioma de destino para calcular una estimación de *P(T)*, y se usan modelos de alineación de complejidad variable para calcular y estimar *P(S | T)*.

Hay un número de problemas asociados a los sistemas convencionales de MT estadística basada en cadenas. En particular, el espacio de búsqueda (todas las posibles cadenas en el idioma de destino) es bastante grande. Sin restringir este espacio de búsqueda, no puede construirse un sistema práctico de MT, porque lleva demasiado tiempo considerar todas las posibles cadenas de traducción. Para abordar esto, muchos sistemas usan una hipótesis simplificadora en cuanto a que las probabilidades del modelo del canal y del modelo del idioma de destino para una cadena entera pueden ser determinadas como el producto de probabilidades de subcadenas dentro de la cadena. Esta hipótesis es solamente válida mientras las dependencias en las cadenas y entre las cadenas estén limitadas a las áreas locales definidas por las subcadenas. Sin embargo, a veces la mejor traducción para un trozo de texto del idioma de origen está condicionada por los elementos de las cadenas del idioma de origen y de destino que están relativamente lejos del elemento a predecir. Dado que las hipótesis simplificadoras hechas en los modelos de MT estadística basada en cadenas están basadas, en gran parte, en la localidad de las cadenas, a veces los elementos condicionantes están lo bastante lejos del elemento a predecir como para que no puedan ser tenidos en cuenta por los modelos.

Por ejemplo, algunos sistemas de MT estadística basada en cadenas usan modelos de n-gramas de cadenas para su modelo lingüístico (LM). Estos modelos de n-gramas son sencillos de entrenar, usar y optimizar. Sin embargo, los modelos de n-gramas tienen algunas limitaciones. Aunque una palabra puede ser predicha con precisión a partir de uno o dos de sus predecesores inmediatos, un buen número de construcciones lingüísticas colocan palabras sumamente predictivas lo suficientemente lejos de las palabras que predicen como para que estén excluidas del alcance del modelo de n-gramas de cadenas. Consideremos las siguientes oraciones activas y pasivas:

1. Juan golpeó la pelota.

2. Las pelotas fueron golpeadas por Lucía.

Los siguientes trigramas ocurren en estas oraciones con las frecuencias indicadas

<P> <P> Juan 1	<P> <P> Las 1
<P> Juan golpeó 1	<P> Las pelotas 1
Juan golpeó la 1	las pelotas fueron 1
golpeó la pelota 1	pelotas fueron golpeadas 1
la pelota <POSTE> 1	fueron golpeadas por 1
	golpeadas por Lucía 1
	por Lucía <POSTE> 1

donde “<P>” es una señal imaginaria al comienzo de una oración que proporciona contexto inicial de oración, y “<POSTE>” es una señal imaginaria al final de una oración. Debería observarse que cada uno de estos trigramas aparece una sola vez, incluso aunque el suceso (el golpe de una pelota) es el mismo en ambos casos.

En otro sistema de MT estadística, una estructura sintáctica en el idioma de origen es correlacionada con una cadena en el idioma de destino. Los modelos basados en la sintaxis tienen varias ventajas sobre los modelos basados en cadenas. En un aspecto, los modelos basados en la sintaxis pueden reducir la magnitud del problema de los datos ralos, normalizando los lemas. En otro aspecto, los modelos basados en la sintaxis pueden tener en cuenta la estructura sintáctica del idioma. Por lo tanto, los sucesos que dependen entre sí están a menudo más cercanos entre sí en un árbol de sintaxis de lo que están en la cadena superficial, porque la distancia hasta un padre común puede ser más corta que la distancia en la cadena.

Sin embargo, incluso en un modelo basado en la sintaxis, quedan inconvenientes: la distancia entre palabras interdependientes puede aún ser demasiado grande para ser capturada por un modelo local; además, conceptos similares son expresados por estructuras distintas (p. ej., voz activa contra voz pasiva) y, por lo tanto, no se modelan juntos. Estos dan como resultado un mal entrenamiento del modelo y malas prestaciones de traducción.

A partir del documento US 2003 / 023422 A1, se conoce un sistema hermético de traducción automática.

Resumen de la invención

La presente invención incluye un procedimiento según la reivindicación 1 y un sistema de traducción automática según la reivindicación 16. Las realizaciones preferidas se revelan en las reivindicaciones dependientes.

Breve descripción de los dibujos

La FIG. 1 ilustra un diagrama de bloques de un entorno informático general en el cual puede ser puesta en práctica la presente invención.

La FIG. 2 ilustra un diagrama de bloques de un dispositivo móvil en el cual puede ser puesta en práctica la presente invención.

Las FIGS. 3 y 4 ilustran ejemplos de formas lógicas.

La FIG. 5 ilustra un diagrama de bloques de una arquitectura de traducción automática de acuerdo a una realización de la presente invención.

La FIG. 6 ilustra una forma lógica ejemplar de destino en el sector de destino de una correlación de transferencias.

La FIG. 7 ilustra un ejemplo de una forma lógica de entrada.

La FIG. 8 ilustra correlaciones ejemplares de transferencias almacenadas en una base de datos de correlaciones de transferencias.

Las FIGS. 9 y 10 ilustran correlaciones ejemplares de transferencias almacenadas en una base de datos de correlaciones de transferencias.

La FIG. 11 ilustra una forma lógica ejemplar de entrada.

La FIG. 12 ilustra una correlación ejemplar de transferencias almacenada en una base de datos de correlaciones de transferencias.

La FIG. 13 es un diagrama de flujo que ilustra un algoritmo de descodificación de acuerdo a la presente invención.

La FIG. 14 ilustra una forma lógica ejemplar de origen con la cual puede ser utilizado el diagrama de flujo de la FIG. 13.

Las FIGS. 15 a 21 ilustran correlaciones ejemplares de transferencias almacenadas en una base de datos de correlaciones de transferencias, con las cuales puede ser utilizado el diagrama de flujo de la FIG. 13.

Descripción detallada de realizaciones ilustrativas

La FIG. 1 ilustra un ejemplo de un entorno adecuado 100 de sistema informático en el cual puede ser implementada la invención. El entorno 100 de sistema informático es solamente un ejemplo de un entorno informático adecuado y no está concebido para sugerir limitación alguna en cuanto al alcance del uso, o la funcionalidad, de la invención. Tampoco debería ser interpretado que el entorno informático 100 tiene dependencia o requisito alguno referido a uno cualquiera, o una combinación, de los componentes ilustrados en el entorno operativo ejemplar 100.

La invención es operativa con otros numerosos entornos o configuraciones de sistema informático de propósito general o de propósito especial. Los ejemplos de sistemas, entornos y / o configuraciones informáticas bien conocidas que pueden ser adecuadas para su uso con la invención incluyen, pero no se limitan a, ordenadores personales, ordenadores servidores, dispositivos de mano o portátiles, sistemas multiprocesadores, sistemas basados en microprocesadores, equipos de sobremesa, electrónica programable de consumo, ordenadores personales en red, miniordenadores, ordenadores centrales, sistemas de telefonía, entornos informáticos distribuidos que incluyen cualquiera de los sistemas o dispositivos anteriores, y similares.

La invención puede ser descrita en el contexto general de las instrucciones ejecutables por ordenador, tales como los módulos de programa que son ejecutados por un ordenador. En general, los módulos de programa incluyen rutinas, programas, objetos, componentes, estructuras de datos, etc., que realizan tareas específicas o implementan tipos específicos de datos abstractos. La invención puede ser puesta en práctica en entornos informáticos distribuidos donde las tareas son realizadas por dispositivos de procesamiento remoto que están enlazados a través de una red de comunicaciones. En un entorno informático distribuido, los módulos de programa están situados en medios de almacenamiento de ordenadores tanto locales como remotos, incluso dispositivos de almacenamiento en memoria.

Con referencia a la FIG. 1, un sistema ejemplar para implementar la invención incluye un dispositivo informático de propósito general en forma de un ordenador 110. Los componentes del ordenador 110 pueden incluir, pero no se limitan a, una unidad 120 de procesamiento, una memoria 130 del sistema y un bus 121 del sistema que acopla

diversos componentes del sistema, incluso la memoria del sistema con la unidad de procesamiento. El bus 121 del sistema puede ser de cualquiera entre diversos tipos de estructuras de bus, incluso un bus de memoria o controlador de memoria, un bus periférico y un bus local que use cualquiera entre una amplia variedad de arquitecturas de bus. A modo de ejemplo, y no de limitación, tales arquitecturas incluyen el bus de Arquitectura Estándar Industrial (ISA), el bus de Arquitectura de Micro Canal (MCA), el bus de ISA Realzada (EISA), el bus local de la Asociación de Estándares Electrónicos de Vídeo (VESA) y el bus de Interconexión de Componentes Periféricos (PCI), también conocido como el bus de Entresuelo.

El ordenador 110 habitualmente incluye una amplia variedad de medios legibles por ordenador. Los medios legibles por ordenador pueden ser medios disponibles cualesquiera a los que pueda acceder el ordenador 110, e incluyen medios tanto volátiles como no volátiles, medios extraíbles y no extraíbles. A modo de ejemplo, y no de limitación, los medios legibles por ordenador pueden comprender medios de almacenamiento de ordenadores y medios de comunicación. Los medios de almacenamiento de ordenadores incluyen medios tanto volátiles como no volátiles, extraíbles y no extraíbles, implementados en cualquier procedimiento o tecnología para el almacenamiento de información, tales como instrucciones legibles por ordenador, estructuras de datos, módulos de programa u otros datos. Los medios de almacenamiento de ordenadores incluyen, pero no se limitan a, memoria RAM, ROM, EEPROM, memoria flash u otra tecnología de memoria, CD-ROM, discos versátiles digitales (DVD) u otro almacenamiento en disco óptico, casetes magnéticos, cinta magnética, almacenamiento en disco magnético u otros dispositivos de almacenamiento magnético, o cualquier otro medio que pueda ser usado para almacenar la información deseada y a la cual pueda acceder el ordenador 110. Los medios de comunicación realizan habitualmente instrucciones legibles por ordenador, estructuras de datos, módulos de programa u otros datos en una señal de datos modulados, tal como una onda portadora u otro mecanismo de transporte, e incluyen cualquier medio de suministro de información. El término “señal de datos modulados” significa una señal que tiene una o más de sus características fijadas o cambiadas de manera tal como para codificar información en la señal. A modo de ejemplo, y no de limitación, los medios de comunicación incluyen medios cableados, tales como una red cableada o conexión de cableado directo, y medios inalámbricos tales como los medios acústicos, de frecuencia de radio, infrarrojos u otros medios inalámbricos. Las combinaciones de cualesquiera de los anteriores también deberían ser incluidas dentro del alcance de los medios legibles por ordenador.

La memoria 130 del sistema incluye medios de almacenamiento de ordenador en forma de memoria volátil y / o no volátil, tal como la memoria de solo lectura (ROM) 131 y la memoria de acceso aleatorio (RAM) 132. Un sistema 133 básico de entrada / salida (BIOS), que contiene las rutinas básicas que ayudan a transferir información entre elementos dentro del ordenador 110, tal como durante el arranque, está almacenado habitualmente en la ROM 131. La RAM 132 contiene habitualmente datos y / o módulos de programa que son inmediatamente accesibles, y / o que están actualmente siendo operados, por la unidad 120 de procesamiento. A modo de ejemplo, y no de limitación, la FIG. 1 ilustra el sistema operativo 134, los programas 135 de aplicación, otros módulos 136 de programa y los datos 137 de programa.

El ordenador 110 también puede incluir otros medios de almacenamiento de ordenador extraíbles o no extraíbles, volátiles o no volátiles. Solamente a modo de ejemplo, la FIG. 1 ilustra un controlador 141 de disco rígido que lee de, o escribe en, medios magnéticos no extraíbles y no volátiles, un controlador 151 de disco magnético que lee de, o escribe en, un disco 152 magnético extraíble y no volátil, y un controlador 155 de disco óptico que lee de, o escribe en, un disco 156 óptico extraíble y no volátil, tal como un CD ROM u otros medios ópticos. Otros medios de almacenamiento de ordenador, extraíbles o no extraíbles, volátiles o no volátiles, que pueden ser usados en el entorno operativo ejemplar incluyen, pero no se limitan a, casetes de cinta magnética, tarjetas de memoria flash, discos versátiles digitales, cinta digital de vídeo, memoria RAM de estado sólido, memoria ROM de estado sólido y similares. El controlador 141 de disco rígido está habitualmente conectado con el bus 121 del sistema a través de una interfaz de memoria no extraíble, tal como la interfaz 140, y el controlador 151 de disco magnético y el controlador 155 de disco óptico están habitualmente conectados con el bus 121 del sistema por una interfaz de memoria extraíble, tal como la interfaz 150.

Los controladores y sus medios asociados de almacenamiento de ordenador, expuestos en lo que antecede e ilustrados en la FIG. 1, proporcionan almacenamiento de instrucciones legibles por ordenador, estructuras de datos, módulos de programa y otros datos para el ordenador 110. En la FIG. 1, por ejemplo, el controlador 141 de disco rígido está ilustrado almacenando el sistema operativo 144, los programas 145 de aplicación, otros módulos 146 de programa y los datos 147 de programa. Obsérvese que estos componentes pueden ser tanto los mismos que, o distintos a, el sistema operativo 134, los programas 135 de aplicación, otros módulos 136 de programa y los datos 137 de programa. Al sistema operativo 144, los programas 145 de aplicación, otros módulos 146 de programa y los datos 147 de programa se dan aquí números distintos, para ilustrar que, como mínimo, son copias distintas.

Un usuario puede ingresar comandos e información en el ordenador 110 a través de dispositivos de entrada tales como un teclado 162, un micrófono 163, un dispositivo puntero 161, tal como un ratón, una bola de rastreo o un panel táctil. Otros dispositivos de entrada (no mostrados) pueden incluir una palanca de juegos, un panel de juegos, una antena parabólica satelital, un escáner o similares. Estos y otros dispositivos de entrada están a menudo conectados con la

unidad 120 de procesamiento a través de una interfaz 160 de entrada de usuario que está acoplada con el bus del sistema, pero pueden estar conectados por otra interfaz y otras estructuras de bus, tales como un puerto paralelo, un puerto de juegos o un bus universal en serie (USB). Un monitor 191, u otro tipo de dispositivo visor, también está conectado con el bus 121 del sistema mediante una interfaz, tal como una interfaz 190 de vídeo. Además del monitor, los ordenadores también pueden incluir otros dispositivos periféricos de salida tales como los altavoces 197 y la impresora 196, que pueden estar conectados a través de una interfaz 195 periférica de salida.

El ordenador 110 es operado en un entorno en red, usando conexiones lógicas con uno o más ordenadores remotos, tal como un ordenador remoto 180. El ordenador remoto 180 puede ser un ordenador personal, un dispositivo de mano, un servidor, un encaminador, un ordenador personal en red, un dispositivo a la par u otro nodo común de red, y habitualmente incluye muchos de, o todos, los elementos descritos anteriormente con respecto al ordenador 110. Las conexiones lógicas ilustradas en la FIG.1 incluyen una red de área local (LAN) 171 y una red de área amplia (WAN) 173, pero también pueden incluir otras redes. Tales entornos de red son comunes en oficinas, redes de ordenadores de ámbito empresarial, intranets e Internet.

Cuando se usa en un entorno de red LAN, el ordenador 110 está conectado con la LAN 171 a través de una interfaz o adaptador 170 de red. Cuando se usa en un entorno de red WAN, el ordenador 110 incluye habitualmente un módem 170 u otro medio para establecer comunicaciones por la WAN 173, tal como Internet. El módem 172, que puede ser interno o externo, puede estar conectado con el bus 121 del sistema mediante la interfaz 160 de entrada de usuario, u otro mecanismo adecuado. En un entorno en red, los módulos de programa ilustrados con respecto al ordenador 110, o partes de los mismos, pueden estar almacenados en el dispositivo de almacenamiento de memoria remota. A modo de ejemplo, y no de limitación, la FIG. 1 ilustra los programas 185 de aplicación remota como residentes en el ordenador remoto 180. Se apreciará que las conexiones de red mostradas son ejemplares, y que pueden usarse otros medios de establecer un enlace de comunicaciones entre los ordenadores.

La FIG. 2 es diagrama de bloques de un dispositivo móvil 200, que es un entorno informático ejemplar. El dispositivo móvil 200 incluye un microprocesador 202, una memoria 204, componentes 206 de entrada / salida (I / O) y una interfaz 208 de comunicación para comunicarse con ordenadores remotos u otros dispositivos móviles. En una disposición, los componentes precitados están acoplados para la comunicación entre sí por un bus 210 adecuado.

La memoria 204 está implementada como memoria electrónica no volátil, tal como memoria de acceso aleatorio (RAM), con un módulo de resguardo de baterías (no mostrado), de modo que la información almacenada en la memoria 204 no se pierda cuando se apaga la energía general para el dispositivo móvil 200. Una parte de la memoria 204 está preferiblemente adjudicada como memoria direccionable para la ejecución de programas, mientras que otra parte de la memoria 204 es usada, preferiblemente, para el almacenamiento, tal como para simular el almacenamiento en un controlador de disco.

La memoria 204 incluye un sistema operativo 212 y programas 214 de aplicación, así como un almacén 216 de objetos. Durante el funcionamiento, el sistema operativo 212 es preferiblemente ejecutado por el procesador 202 desde la memoria 204. El sistema operativo 212, en una disposición preferida, es el sistema operativo de marca WINDOWS® CE, disponible comercialmente en la Corporación Microsoft. El sistema operativo 212 está preferiblemente diseñado para dispositivos móviles, e implementa características de base de datos que pueden ser utilizadas por las aplicaciones 214, a través de un conjunto de interfaces y procedimientos expuestos de programación de aplicaciones. Los objetos en el almacén 216 de objetos son mantenidos por las aplicaciones 214 y el sistema operativo 212, al menos parcialmente, en respuesta a llamadas a las interfaces y procedimientos expuestos de programación de aplicaciones.

La interfaz 208 de comunicación representa a numerosos dispositivos y tecnologías que permiten al dispositivo móvil 200 enviar y recibir información. Los dispositivos incluyen módems cableados e inalámbricos, receptores por satélite y sintonizadores de emisión, por nombrar unos pocos. El dispositivo móvil 200 también puede ser conectado directamente con un ordenador para intercambiar datos con el mismo. En tales casos, la interfaz 208 de comunicación puede ser un transceptor infrarrojo o una conexión de comunicación en serie o en paralelo, todos los cuales son capaces de transmitir información por flujo.

Los componentes 206 de entrada / salida incluyen una amplia variedad de dispositivos de entrada, tales como una pantalla sensible al tacto, botones, rodillos y un micrófono, así como una amplia variedad de dispositivos de salida que incluyen un generador de audio, un dispositivo vibrador y un visor. Los dispositivos enumerados anteriormente son a modo de ejemplo y no necesariamente deben estar todos presentes en el dispositivo móvil 200. Además, otros dispositivos de entrada / salida pueden estar adosados a, o hallados en, el dispositivo móvil 200, dentro del alcance de la presente invención.

Formas lógicas

Antes de exponer la presente invención en mayor detalle, puede ser útil una breve exposición de una forma lógica. Una exposición completa y detallada de las formas lógicas, y de los sistemas y procedimientos para generarlas, puede

hallarse en la Patente Estadounidense Nº 5.966.686 de Heidom et al., publicada el 12 de octubre de 1999 y titulada PROCEDIMIENTO Y SISTEMA PARA CALCULAR FORMAS LÓGICAS SEMÁNTICAS A PARTIR DE ÁRBOLES DE SINTAXIS. En breve, sin embargo, las formas lógicas son generadas realizando un análisis morfológico sobre un texto de entrada, para producir análisis de estructuras de frases convencionales, aumentadas con relaciones gramaticales. Los análisis sintácticos se someten a un procesamiento adicional a fin de obtener formas lógicas, que son estructuras de datos que describen dependencias etiquetadas entre palabras del contenido en la entrada textual.

En general, una forma lógica es una estructura de datos de relaciones lógicas conectadas, que representa una única entrada, tal como una oración o parte de la misma. La forma lógica, mínimamente, consiste en una relación lógica y refleja relaciones estructurales (es decir, relaciones sintácticas y semánticas), en particular, una o más relaciones de argumentos, o adjuntas, entre palabras importantes en una cadena de entrada.

Las formas lógicas pueden normalizar ciertas alternancias sintácticas (p. ej., activo / pasivo) y resolver anáforas entre oraciones y dependencias a larga distancia. Por ejemplo, las FIGS. 3 y 4 ilustran formas lógicas o gráficos 300 y 400 de dependencia para las oraciones activas y pasivas dadas como ejemplos en la sección de Trasfondo, para ayudar en la comprensión de los elementos de las formas lógicas. Sin embargo, como será apreciado por los expertos en la técnica, cuando son almacenadas en un medio legible por ordenador, las formas lógicas pueden no ser inmediatamente entendidas como representantes de un gráfico. Las FIGS. 3 y 4 ilustran importantes generalizaciones que la cadena superficial y los modelos de sintaxis, según lo descrito en la sección de Trasfondo, no pueden capturar.

Para ver por qué los gráficos de dependencia podrían proporcionar un mejor modelo lingüístico que los modelos de n-gramas basados en cadenas o los árboles de sintaxis, considérense las siguientes oraciones:

1. Juan golpeó la pelota.

2. Las pelotas fueron golpeadas por Lucía.

Un modelo de trigramas basado en cadenas superficiales generaría los siguientes totales:

	<P> <P> Las 1
<P> <P> Juan 1	<P> Las pelotas 1
<P> Juan golpeó 1	las pelotas fueron 1
Juan golpeó la 1	pelotas fueron golpeadas 1
golpeó la pelota 1	fueron golpeadas por 1
la pelota <POSTE> 1	golpeadas por Lucía 1
	por Lucía <POSTE> 1

Obsérvese que cada uno de estos trigramas ocurre solamente una vez, incluso aunque el suceso (el golpe de una pelota) es el mismo en ambos casos.

Si miramos a los árboles de sintaxis, obtenemos una imagen levemente distinta. Específicamente, para las oraciones anteriores, se producirían los siguientes árboles de sintaxis.

Juan golpeó la pelota.

```
Decl1
    | _________ NP1 _____ SUSTANTIVO1    “Juan”
    | _________ VERBO1        “golpeó”
    | _________ NP2 _____ DETP1 ____ ADJ1*    “la”
    |                   | _____ SUSTANTIVO2    “pelota”
    | _________ CAR1   “.”
```

Las pelotas fueron golpeadas por Lucía.

```
Decl1
    | _________ NP1 _____ DETP1 _____ ADJ1*    “Las”
```

| | _____ SUSTANTIVO1 “pelotas”

| _____ AUX1 _____ VERBO1 “fueron”

| _____ VERBO2* “golpeadas”

| _____ PP1 _____ PP2 _____ PREP1* “por”

| | _____ SUSTANTIVO2* “Lucía”

| _____ CAR1 “.”

Obsérvese que aquí también, el golpe de la pelota está divido entre dos cubos distintos (un conjunto de reglas para la voz activa y otro para la voz pasiva) y, por tanto, el sistema no lograría aprender una generalización útil.

Las FIGS. 3 y 4 ilustran las formas lógicas 300 y 400. Las LF 300 y 400 incluyen los nodos paternos 302 y 402, los nodos filiales 304, 308, 404 y 408 y los nodos 306, 307, 406 y 407 de relaciones semánticas. Los nodos 306 y 406 de relaciones semánticas operan para conectar los nodos filiales 304, 308, 404 y 408 con los nodos paternos 302 y 402, y explican la relación semántica entre los nodos paternos y filiales.

Los nodos paternos 302 y 402 contienen formas o lemas de palabras. Por ejemplo, el lema en los nodos paternos 302 y 402 es la palabra “golpe”. Los nodos filiales 304, 308, 404 y 408 también contienen formas o lemas de palabras. Los nodos 306 y 406 de relaciones semánticas ilustran que los nodos filiales 304 y 404 son sujetos profundos, y los nodos 307 y 407 de relaciones semánticas indican que los nodos filiales 308 y 408 son objetos profundos de los nodos paternos 302 y 402. Además, las LF 300 y 400 también incluyen características binarias (o “bits”) adosadas a cada lema en cada nodo. Por ejemplo, las características binarias están adosadas a cada lema de las LF 300 y 400, y se ilustran en paréntesis. Las características binarias describen las propiedades sintácticas de un lema. Por ejemplo, la forma de palabras en el nodo 302 incluye bits que describen la palabra “golpe” como un tiempo pasado y como una preposición.

A diferencia de las cadenas y los árboles de sintaxis, las formas lógicas 300 y 400 incluyen tanto la construcción en voz activa como la construcción en voz pasiva en la misma estructura de gráfico. Las estructuras de LF pueden ser desgraficadas para producir un árbol, de modo que ningún nodo pueda tener dos padres. Las formas lógicas 300 y 400 son representaciones lógicas jerárquicas de las correspondientes oraciones (o fragmentos de oración). Cada nodo depende de todos sus ancestros. Con el fin de construir un modelo lingüístico (de destino), la estructura local se modela en base a la aproximación de que cada nodo filial depende solamente de su padre (o de n-1 ancestros, para un modelo de LF de n-gramas).

En una disposición ilustrativa, el código específico que construye formas lógicas a partir de análisis sintácticos es compartido entre los diversos idiomas de origen y de destino sobre los cuales opera el sistema de traducción automática. La arquitectura compartida simplifica en gran medida la tarea de alinear los segmentos de formas lógicas provenientes de distintos idiomas, ya que las construcciones superficialmente distintas en dos idiomas se colapsan frecuentemente sobre representaciones de formas lógicas similares o idénticas.

Traducción Automática

La FIG. 5 es un diagrama en bloques de una arquitectura ejemplar de un sistema 500 de traducción automática de acuerdo a la presente invención. El sistema 500 incluye los componentes 504 y 506 de análisis sintáctico, el componente estadístico 508 de aprendizaje de asociaciones de palabras, el componente 510 de alineación de formas lógicas, el componente 512 de construcción de la base de conocimiento léxico (LKB), el diccionario bilingüe 514, la lista 520 de asociaciones y la base 518 de datos de correlaciones de transferencias. Durante el tiempo de ejecución de la traducción, el sistema 500 utiliza el componente 522 de análisis sintáctico, el componente 524 de búsqueda, el componente 554 de descodificación, el componente 526 de transferencia y el componente 528 de generación.

Según la invención, se usa un corpus bilingüe para entrenar el sistema. El corpus bilingüe (o “bitexto”) incluye oraciones traducidas alineadas (p. ej., oraciones en un idioma de origen o de destino, tal como el inglés, en una correspondencia de 1 a 1 con sus traducciones creadas por seres humanos en el otro idioma de origen o de destino, tal como el castellano). Durante el entrenamiento, se proporcionan oraciones provenientes del corpus bilingüe alineado al sistema 500, como las oraciones 530 de origen (las oraciones a traducir), y como las oraciones 532 de destino (la traducción de las oraciones de origen). Los componentes 504 y 506 de análisis sintáctico analizan sintácticamente las oraciones de origen y las oraciones de destino, provenientes del corpus bilingüe alineado, para producir las formas lógicas 534 de origen y las formas lógicas 536 de destino.

Durante el análisis sintáctico, las palabras en las oraciones son convertidas en formas o lemas de palabras normalizadas, y pueden ser proporcionadas al componente estadístico 508 de aprendizaje de asociaciones de

palabras. Ambas asociaciones, de palabras únicas y de palabras múltiples, son hipotetizadas iterativamente y calificadas por el componente 508 de aprendizaje, hasta que se obtenga un conjunto fiable de cada una. El componente estadístico 508 de aprendizaje de asociaciones de palabras emite los pares 538 de traducciones de palabras aprendidas.

Los pares de palabras son añadidos a una lista 520 de asociaciones, que actúa como un diccionario bilingüe actualizado.

Los pares 538 de palabras, junto con las formas lógicas 534 de origen y las formas lógicas 536 de destino, son proporcionados al componente 510 de alineación de formas lógicas. En breve, el componente 510 establece primero correspondencias tentativas, respectivamente, entre los nodos en las formas lógicas 530 y 536, de origen y de destino. Esto se hace usando pares de traducciones provenientes de un léxico bilingüe (p. ej., un diccionario bilingüe) 514, que puede ser aumentado con los pares 538 de palabras provenientes del componente estadístico 508 de aprendizaje de asociaciones de palabras. Después de establecer las posibles correspondencias, el componente 510 de alineación alinea los nodos de formas lógicas según consideraciones tanto léxicas como estructurales, y crea las correlaciones 542 de transferencias de palabras y / o de formas lógicas.

Básicamente, el componente 510 de alineación traza enlaces entre formas lógicas usando la información 518 del diccionario bilingüe y los pares 538 de palabras. Las correlaciones de transferencias son filtradas optativamente, en base a una frecuencia con la cual se hallan en las formas lógicas 534 y 536, de origen y de destino, y son proporcionadas a un componente 512 de construcción de base de conocimiento léxico.

Si bien el filtrado es optativo, en un ejemplo, si la correlación de transferencia no es vista al menos dos veces en los datos de entrenamiento, no es usada para construir la base 518 de datos de correlaciones de transferencias, aunque cualquier otra frecuencia deseada puede ser asimismo usada como un filtro. También debería observarse que pueden ser usadas otras técnicas de filtrado, que no sean la frecuencia de aparición. Por ejemplo, las correlaciones de transferencias pueden ser filtradas en base a si están o no formadas a partir de análisis sintácticos completos de las oraciones de origen, y en base a si las formas lógicas usadas para crear las correlaciones de transferencias están o no completamente alineadas.

El componente 512 construye la base 518 de datos de correlaciones de transferencias, que contiene las correlaciones de transferencias que enlazan básicamente las palabras y / o formas lógicas en un idioma, con las palabras y / o formas lógicas en el segundo idioma. Con la base 518 de datos de correlaciones de transferencias creada de tal modo, el sistema 500 está ahora configurado para traducciones en tiempo de ejecución.

Durante el tiempo de ejecución de la traducción, un texto 550 de origen, a traducir, es proporcionado al componente 522 de análisis sintáctico. El componente 522 de análisis sintáctico recibe el texto 550 de origen y crea una forma lógica 552 de origen en base a la entrada del texto de origen.

La forma lógica 552 de origen es proporcionada al componente 524 de búsqueda. El componente 524 de búsqueda intenta buscar en la base 518 de datos de correlaciones de transferencias, a fin de obtener las correlaciones de transferencias que cubran toda, o partes de, la forma lógica 552 de origen. Pueden hallarse correlaciones múltiples para uno o más de los nodos en la forma lógica 552 de origen.

Después de que se hallan un conjunto de posibles correlaciones de transferencias y un conjunto de posibles combinaciones de correlaciones de transferencias, el componente 554 de descodificación puntúa cada combinación de correlaciones de transferencia, usando una pluralidad de modelos. Según una disposición, las correlaciones de transferencias individuales y las combinaciones de correlaciones de transferencias son calificadas con un puntaje linealmente interpolado que será explicado en mayor detalle más adelante. Después de que se generan los puntajes, el componente 554 de descodificación escoge y almacena la mejor combinación de correlaciones de transferencias.

El componente 526 de transferencias recibe la mejor combinación de correlaciones candidatas desde el componente 554 de descodificación y construye una forma lógica 556 de destino que formará la base de la traducción de destino. Esto se hace combinando los nodos de las correlaciones de transferencias seleccionadas.

En los casos en que no sea hallada ninguna correlación de transferencias aplicable por el componente 524 de búsqueda, los nodos en la forma lógica 552 de origen y sus relaciones son sencillamente copiados a la forma lógica 556 de destino. Las traducciones por omisión de palabras únicas aún pueden ser halladas en la base 518 de datos de correlaciones de transferencias para estos nodos, y pueden ser insertadas en la forma lógica 556 de destino. Sin embargo, si no se halla ninguna, las traducciones, de forma ilustrativa, pueden ser obtenidas de la lista 520 de asociaciones, que fuera usada durante la alineación.

El componente 528 de generación, ilustrativamente, es un componente de generación basado en reglas e independiente de las aplicaciones, que correlaciona la forma lógica 556 de destino con el texto de destino (o cadena de destino) 558. El componente 528 de generación, alternativamente, puede utilizar un enfoque aprendido

automáticamente para la realización de oraciones. El componente 528 de generación, de manera ilustrativa, puede no tener ninguna información con respecto al idioma de origen de las formas lógicas de entrada, y funciona exclusivamente con información que le pasa el componente 526 de transferencia. El componente 528 de generación, también de manera ilustrativa, usa esta información conjuntamente con un diccionario monolingüe (p. ej., para el idioma de destino), para producir el texto 558 de destino. Un componente 528 genérico de generación es, por tanto, suficiente para cada idioma.

Traducción automática estadística de formas lógicas

Con respecto a la siguiente exposición, los expertos en la técnica deberían reconocer que todos los gráficos de dependencia, formas lógicas, estructuras semánticas, relaciones semánticas y representaciones semánticas se refieren a, y describen, la forma lógica de entrada según lo proporcionado durante el tiempo de ejecución. Además, los expertos en la técnica deberían reconocer que las correlaciones de transferencia, o correlaciones, se refieren a aquellas correlaciones formadas durante el entrenamiento.

La siguiente ecuación, reminiscente de la ecuación 2, es una vista de alto nivel de una disposición de la presente invención:

Ecuación 3

$$T = \arg\max_{T'} P(S \mid T') \times P(T')$$

donde T' está restringido a ser una forma lógica en el idioma de destino, $P(S \mid T')$ es la probabilidad de la forma lógica S de origen, dada una forma lógica T' del idioma de destino, y $P(T')$ es la probabilidad de la forma lógica T' del idioma de destino, también escrita como $P_{\mu T}(T')$, donde μ_T es un modelo del idioma de destino. La ecuación anterior es equivalente a la:

Ecuación 4

$$T = \arg\max_{T'} [\log P(S \mid T') + \log P_{\mu T}(T')]$$

Una disposición de la presente invención aproxima $P(S \mid T')$ incorporando varias fuentes de conocimiento: un modelo de canal o un modelo de traducción $P_{\mu C}(S, T')$, un modelo de fertilidad $P_{\mu F}(S, T')$, una fuente de información sobre el tamaño de la correlación, $Puntaje_{\mu S}(S, T')$, y una fuente de información de correlación (o "rango") de características binarias $Puntaje_{\mu B}(S, T')$. Al incorporar estas fuentes de conocimiento, $P(S \mid T')$ se aproxima de la siguiente manera:

Ecuación 5

$$\log P(S \mid T') \approx \log P_{\mu C}(S, T') + \log P_{\mu F}(S, T') + Puntaje_{\mu S}(S, T') + Puntaje_{\mu B}(S, T')$$

Por tanto,

Ecuación 6

$$T \approx \arg\max_{T'} \left[\log P_{\mu C}(S, T') + \log P_{\mu F}(S, T') + Puntaje_{\mu S}(S, T') + Puntaje_{\mu B}(S, T') + \log P_{\mu T}(T') \right]$$

La contribución relativa de cada puntaje, o log-probabilidad, está ponderada por un factor (λ_T, λ_C,λ_F,λ_S y λ_B) y la aproximación interpolada lineal resultante es la:

Ecuación 7

$$T \approx \arg\max_{T'} \left[\lambda_C \cdot \log P_{\mu C}(S, T') + \lambda_F \cdot \log P_{\mu F}(S, T') + \lambda_S \cdot Puntaje_{\mu S}(S, T') + \lambda_B \cdot Puntaje_{\mu B}(S, T') + \lambda_T \cdot \log P_{\mu T}(T') \right]$$

En la práctica, esta disposición no puntúa una forma lógica entera de origen, y una forma lógica del idioma de destino, todo a la vez. En cambio, la búsqueda (representada en las ecuaciones anteriores por el "argmax") construye una forma lógica del idioma de destino, una correlación de traducción a la vez. Al hacerlo, emplea un puntaje para cada correlación. La puntuación total, linealmente interpolada, para una correlación m de transferencia está representada por

la:

Ecuación 8

$$PUNTAJE\ (m) = \log P(m) = \lambda_T \cdot \log P_{\mu T}(m) + \lambda_C \cdot \log P_{\mu C}(m)$$

$$+ \lambda_F \cdot \log P_{\mu F}(m) + \lambda_S \cdot Puntaje_{\mu S}(m) + \lambda_B \cdot \log Puntaje_{\mu B}(m)$$

Como en la aproximación completa, y según lo indicado anteriormente, cada puntaje o probabilidad de fuente de información está ponderado por un factor. Estos factores (λ_T, λ_C,λ_F,λ_S y λ_B) o pesos son entrenados usando el algoritmo de Powell para maximizar el puntaje BLEU en la salida del sistema. El algoritmo de Powell es conocido en la técnica. Un ejemplo de este algoritmo está descrito en un artículo de Powell titulado “Un procedimiento eficaz para hallar el mínimo de una función de varias variables sin calcular derivadas” (Computer Journal, 7; 155-162). El puntaje BLEU también es conocido en la técnica y está descrito en un artículo de Papineni, Roukos, Ward y Zhu titulado “Bleu: un procedimiento para la evaluación automática de la traducción automática”, 2001, Informe Técnico de IBM RC22176 (W0109-022), División de Investigación de IBM, Centro de Investigación Thomas J. Watson.

Modelos

Según una disposición, el modelo del idioma de destino es un modelo de n-gramas que proporciona la probabilidad de un nodo en un gráfico de dependencia de destino, dada una secuencia de n-1 nodos y relaciones precedentes. La FIG. 6 ilustra un gráfico ejemplar 600 de dependencia de destino, que puede ser hallado en el sector de destino en la base 518 de datos de correlaciones de transferencias. En la FIG. 6, los nodos A, B, C y D contienen formas de palabras. Los nodos B, C y D son nodos filiales y el nodo A es un nodo padre o nodo raíz. Los Nodos R1 y R2 son nodos de relaciones semánticas.

Usando este modelo de n-gramas, la probabilidad de todo el gráfico τ 600 de dependencia de destino, dado el modelo del idioma de destino, es igual al producto de las probabilidades de n-gramas de cada uno de los nodos. De este modo, la probabilidad del gráfico 600 de dependencia del destino, dado el modelo del idioma de destino, está representada por la siguiente fórmula:

Ecuación 9

$$P_{\mu_T}(\tau) = \prod_i P_{\mu_T}\left(c_i \mid c_{i-1} .. c_{i-(n-1)}\right)$$

donde *i* es un índice sobre todos los nodos en la forma lógica τ. Para cada nodo c_i, el puntaje según el modelo del idioma de destino es la probabilidad de *c*, dados sus n-1 ancestros más cercanos, c_{i-1} a $c_{i-(n-1)}$. Por ejemplo, la probabilidad de la forma lógica 600 de destino, según un modelo de trigramas de esta clase, sería la:

Ecuación 10

$$P_{\mu T}(\tau) = P(A \mid RAÍZ) \cdot P(R1 \mid RAÍZ, A) \cdot P(R2 \mid RAÍZ, A)$$

$$\cdot P(B \mid A, R1) \cdot P(C \mid A, R2) \cdot P(D \mid A, R2)$$

$$\cdot P(HOJA \mid R1, B) \cdot P(HOJA \mid R2, C) \cdot P(HOJA \mid R2, D)$$

En este modelo de trigramas, las relaciones semánticas (R1, R2) son tratadas como entidades de primera clase, de modo que el modelo de destino se simplifique a fin de que sean innecesarios modelos distintos para lemas y relaciones semánticas. El modelo de destino se poda quitando los n-gramas que aparecen infrecuentemente y se allana usando el descuento absoluto interpolado, que es conocido en la técnica y está descrito en un artículo de Ney, Essen y Kneser titulado “Sobre la estructuración de dependencias probabilísticas en la modelación estocástica de idiomas”, 1994 (Habla e Idioma por Ordenador, 8:1-38).

El modelo de canal predice la probabilidad de la forma lógica de origen dada por la forma lógica de destino *P (S | T)*. En una disposición, definimos una cobertura *M* de correlación de transferencia, para una forma lógica *S* de origen dada y una forma lógica *T* de destino, como un conjunto de correlaciones de transferencias de *S* a *T* (indicado como *M:S->T*). La probabilidad de la forma lógica *S* de origen dada la forma lógica *T* de destino es estimada por la:

Ecuación 11

$$P(S \mid T) = \sum_{M_i : S \to T} \prod_{m \in M_i} P_{\mu_C}(m)$$

donde *i* varía (potencialmente) sobre todas las coberturas $M_i:S->T$ de correlaciones de transferencias, y la

Ecuación 12

$$P_{\mu C}(m) = \frac{total(m_S, m_T)}{total(m_T)}$$

define la probabilidad de una correlación *m* según el modelo μC de canal. La expresión $total(m_S, m_T)$ es el número de veces que la estructura en el sector de origen de la correlación *m* fue correlacionada con la estructura en el sector de destino de la correlación *m* en un conjunto de datos de entrenamiento, y $total(m_T)$ es el número de veces que la estructura en el sector de destino de la correlación *m* fue hallada como el sector de destino de cualquier correlación en los datos de entrenamiento.

En otras palabras, la probabilidad, según el modelo μC de canal de una correlación *m* de transferencia, es estimada dividiendo cuántas veces el sector de origen de una correlación de transferencia fue encontrado con el sector de destino de una correlación de transferencia, $total(m_S, m_T)$ (en un bitexto de forma lógica), entre cuántas veces el sector de destino de esa correlación de transferencia fue encontrado, $total(m_T)$.

El modelo de canal también usa correlaciones solapadas de transferencias. Así, la probabilidad calculada en la Ecuación 12 es la probabilidad no normalizada. La probabilidad no normalizada puede ser normalizada de modo que el modelo de canal no favorezca ciertas correlaciones que tienen más solapamiento que otras.

La FIG. 7 ilustra una LF 700 de entrada y la FIG. 8 ilustra las correlaciones 800 de transferencias halladas en la base 518 de datos de correlaciones de transferencias. Las correlaciones 800 de transferencias incluyen las correlaciones 802, 803, 804, 806 y 808. Para facilitar la ilustración, solamente los nodos para los lemas en las correlaciones de transferencias se muestran en las figuras de la presente solicitud, y los nodos para las relaciones semánticas no se muestran. Debería observarse que las correlaciones de transferencias incluyen nodos adicionales para las relaciones semánticas. Sin embargo, dado que estos nodos son tratados de la misma manera que los nodos para los lemas según la presente invención, las figuras y la exposición más adelante están limitadas a exponer solamente los nodos de lemas.

Las correlaciones 800 pueden ser combinadas en un cierto número de formas distintas para cubrir todos los nodos de la LF 700 de origen. Por ejemplo, la correlación 802 puede ser combinada con la correlación 808, la correlación 803 puede ser combinada con la correlación 806 y las correlaciones 804, 806 y 808 pueden ser combinadas entre sí. Cada una de estas combinaciones no está solapada, porque cada nodo en la LF 700 de origen está cubierto solamente por una única correlación de transferencia. Sin embargo, otra manera de combinar las correlaciones 800 para cubrir todas las LF 700 de origen es combinar la correlación 802 con la correlación 803. Esto forma una correlación solapada, porque el nodo A de origen está cubierto por ambas correlaciones 802 y 803.

Para prevenir que el modelo de canal favorezca las correlaciones solapadas, se normaliza el modelo de canal. La probabilidad normalizada $P^N_{\mu C}(m)$ P_N para una correlación *m*, dado el modelo μC de canal, se calcula con la:

Ecuación 13

$$P^N_{\mu C}(m) = P_{\mu C}(m)^{\frac{nuevos}{totales}}$$

donde “nuevos” es el número de constituyentes o nodos no cubiertos previamente en el gráfico de entrada, “totales” es el número total de constituyentes en las correlaciones de transferencias del sector de origen y $P_{\mu C}(m)$ es la probabilidad no normalizada de una correlación de transferencia de acuerdo al modelo de canal, según lo definido anteriormente. De este modo, la Ecuación 13 ilustra que la probabilidad normalizada de una correlación de transferencia, según el modelo de canal, es igual a la probabilidad no normalizada de la correlación de transferencia según el modelo de canal que tiene el exponente ilustrado.

Según lo anteriormente expuesto, en algunos casos, un nodo en la LF de entrada no es hallado en los datos de entrenamiento, porque no pudo ser analizado sintácticamente, sencillamente no apareció, o bien fue eliminado por filtrado de la base 518 de datos de correlaciones de transferencias, porque su frecuencia estaba por debajo de un umbral predeterminado (habitualmente, uno). En estos casos, se inserta un nodo por omisión, que se forma usando una traducción de palabra única a partir de un diccionario. La probabilidad para esta correlación por omisión puede ser determinada usando el modelo de IBM, 1 entrenado en un texto bilingüe. Debido a que esta probabilidad se obtiene de manera distinta a la de las otras probabilidades de canal, la probabilidad de la correlación por omisión es ajustada con un peso (λ_L) antes de ser combinada con las otras probabilidades de canal. Además, si no puede determinarse ninguna probabilidad para la correlación por omisión, se usa un valor por omisión de (λ_A). Los pesos λ_L y λ_A, como el resto de

los pesos usados en el sistema de traducción, son entrenados usando el algoritmo de Powell para maximizar el puntaje BLEU para un conjunto de oraciones de entrenamiento. Debería observarse que estos parámetros son distintos al peso (λ_C) asociado al modelo de canal al calcular el puntaje final para una correlación de transferencia, según lo ilustrado en la Ecuación 8.

El modelo de canal opera de manera distinta al modelo del idioma de destino. Específicamente, el modelo de canal promueve la precisión en la traducción, mientras que el modelo del idioma de destino promueve la fluidez en el idioma de destino, sin consideración del idioma de origen.

El modelo del idioma de destino padece un inconveniente, en cuanto a que prefiere los gráficos más pequeños a los más grandes. Como resultado, el modelo del idioma de destino favorece las correlaciones que borran un nodo en la estructura de destino, ante las correlaciones que mantienen el mismo número de nodos en las estructuras de origen y de destino, o que añaden un nodo en la estructura de destino. Por ejemplo, las FIGS. 9 y 10 ilustran que la base 518 de datos de entrenamiento contiene las correlaciones 900 y 1000 de transferencia. La correlación 900 de transferencia ilustra que hay un nodo menos en el sector de destino que en el sector de origen. La correlación 1000 de transferencia ilustra que hay el mismo número de nodos en el sector de origen y en el sector de destino de la correlación. El modelo del idioma de destino puntuará a la correlación 900 más alto que a la correlación 1000, porque hay menos probabilidades en el producto para el fragmento de LF de destino resultante de la correlación 900 que en el fragmento resultante de la correlación 1000.

El modelo de fertilidad ayuda a vencer este problema, proporcionando un puntaje basado en el número de veces que son borrados nodos en las correlaciones en los datos de entrenamiento. Si los nodos son raramente borrados en los datos de entrenamiento, el modelo de fertilidad proporcionará un mayor puntaje para las correlaciones que no tengan borrados.

El modelo de fertilidad se calcula revisando los datos de entrenamiento y contando, para cada nodo en el sector de origen de la correlación de transferencia, con cuánta frecuencia hay un nodo correspondiente en el sector de destino de las correlaciones de transferencia. Para evitar problemas de datos ralos, el recuento para los lemas se agrupan entre sí por las partes del habla, mientras que el recuento para las relaciones semánticas (cuyo número es aproximadamente igual al número de las partes del habla) se mantienen por separado. Estas frecuencias se usan luego para estimar la probabilidad de que ocurra un borrado.

Para cada etiqueta x de una parte del habla o una relación semántica, una entrada en una tabla de fertilidad está representada por la:

Ecuación 14

$$F[x] = \frac{\text{recuento } (x \in m_T,\ x \in m_S)}{\text{recuento } (x \in m_S)}$$

En otras palabras, la entrada de la tabla de fertilidad para el miembro x se llena calculando la razón entre el número de veces que x fue encontrado tanto en la estructura de destino como en la estructura de origen, y el número de veces que el miembro x fue encontrado en la estructura de origen. Por lo tanto, la probabilidad de una correlación *m* de transferencia, según el modelo de fertilidad μ_F, se calcula con la:

Ecuación 15

$$P_{\mu_F}(m) = \prod_{c \in m_t} f(c)$$

donde

$$f(c) = \begin{cases} F[x] & \text{si } \exists\ c_t \in m_t : c_t \text{ corresponde a } c_s \\ 1 - F[x] & \text{en caso contrario} \end{cases}$$

En otras palabras, si un nodo del destino corresponde a un nodo del origen, entonces f(c) es la entrada F[x] de la tabla de fertilidad.

En caso contrario, f(c) es 1-F[x].

La próxima fuente de información es el puntaje de tamaños de correlación, que tiene en cuenta el número de nodos en el sector de origen de las correlaciones de transferencias. Esta fuente de información asigna un puntaje calculado por

la:

Ecuación 16

$$Puntaje_{\mu S}(m) = |m|$$

En efecto, el puntaje del tamaño da preferencia a las mayores correlaciones, sobre la hipótesis de que es probable que las correlaciones con más información de contexto sean mejores que las correlaciones con menos información de contexto. Con referencia a las FIGS. 9 y 10, la correlación 900 de transferencia recibiría un puntaje de dos, porque hay dos nodos en el sector de origen. La correlación 1000 de transferencia también recibiría un puntaje de dos, porque hay dos nodos en el sector de origen.

La fuente de información de las características binarias (o bits) tiene en cuenta el número de características binarias (bits) que coinciden entre el gráfico de dependencia de entrada y el sector de origen de la correlación de transferencia. El origen de las características binarias proporciona un puntaje de clasificación que es la suma de los bits de entrada en el gráfico de dependencia de origen que coinciden con los bits en el sector de origen de la correlación de transferencia. La FIG. 11 ilustra una LF 1100 de entrada y la FIG. 12 ilustra una correlación 1200 de transferencia almacenada en la base 518 de datos de correlaciones de transferencia. El nodo A en la LF 1100 de entrada especifica que el lema del nodo A tiene un bit pasivo y un bit singular. El nodo A del sector de origen de la correlación 1200 de transferencia especifica que el lema del nodo A tiene un bit singular. Por lo tanto, el puntaje de clasificación de la correlación 1200 de transferencia es uno, porque tanto el nodo A de la LF 1100 de entrada como el nodo A del sector de origen de la correlación 1200 tienen un bit singular coincidente.

La FIG. 13 es un diagrama 1300 de flujo que ilustra el algoritmo de descodificación, según lo implementado por el componente 554 de descodificación de la FIG. 5. El componente 554 de descodificación selecciona y puntúa conjuntos de correlaciones de transferencia, de acuerdo a la presente invención. El componente 554 de descodificación usa una búsqueda de arriba hacia abajo, con memorización, para hallar la combinación más probable de correlaciones provenientes del conjunto de correlaciones de transferencia hallados por el componente 524 de búsqueda. La FIG. 14 ilustra la LF 1400 de origen, según un ejemplo de la presente invención. Las FIGS. 15 a 21 ilustran un conjunto ejemplar de correlaciones de transferencia que fueron halladas por el componente 524 de búsqueda, y que se refieren a la LF 1400 de origen.

El componente 554 de descodificación de la FIG. 5 comienza seleccionando el nodo supremo de la LF 1400 de origen, según se muestra en el bloque 1302. En la FIG. 14, el nodo supremo es el nodo A. Después de seleccionar el nodo A, el componente 554 de descodificación pasa al bloque 1304 y determina si la mejor correlación para este nodo, en este contexto, ha sido identificada antes. En este ejemplo, no se ha puntuado ninguna correlación para el nodo A.

El proceso continúa luego en la etapa 1306, donde se selecciona una correlación de transferencia entre el conjunto de correlaciones halladas por el componente 524 de búsqueda que tiene un sector de origen que cubre el nodo seleccionado. Por ejemplo, el componente 554 de descodificación selecciona la correlación 1500 de transferencia de la FIG. 5.

En el bloque 1308, el componente 554 de descodificación determina si hay algún nodo en la LF 1400 de origen que no esté cubierto por la correlación seleccionada y que se extienda directamente desde la correlación seleccionada. En el ejemplo anterior, la correlación 1500 solamente cubre los nodos A y B. De tal modo, el nodo filial C no está cubierto por la correlación seleccionada, pero se extiende directamente desde el nodo A, que está cubierto por la correlación seleccionada. Si hay un nodo filial no cubierto en la etapa 1308, el proceso continúa en el bloque 1310.

En el bloque 1310, el componente 554 de descodificación selecciona el nodo filial C y regresa al bloque 1304. En el bloque 1304, el componente 554 de descodificación determina si ha sido ya identificada una correlación óptima para el nodo seleccionado en el contexto seleccionado. En particular, para un modelo de destino de trigramas, “el contexto seleccionado” consiste en los n-1 ancestros del sector de destino del nodo C (en este caso, <PRE_RAÍZ, PRE_RAÍZ, A’>). Para el nodo C, no ha sido identificada la mejor correlación, por lo que el proceso continúa en la etapa 1306, donde el componente 554 de descodificación selecciona una correlación de transferencia entre el conjunto de correlaciones de transferencia que cubren el nodo filial C. Por ejemplo, el componente 554 de descodificación puede seleccionar la correlación 1600 de transferencia ilustrada en la FIG. 16. Después de la etapa 1306, el componente 554 de descodificación avanza al bloque 1308 y decide si hay algún nodo filial no cubierto que se extienda desde un nodo cubierto por la correlación. En el ejemplo anterior, los nodos E y F son nodos filiales no cubiertos que se extienden desde los nodos cubiertos por la correlación 1600. En base al descubrimiento de nodos filiales no cubiertos, el componente 554 de descodificación avanza al bloque 1310.

En el bloque 1310, el componente 554 de descodificación selecciona uno de los nodos filiales no cubiertos, por ejemplo, el nodo E, y regresa al bloque 1304. En el bloque 1304, el componente 554 de descodificación determina que no ha sido determinada la mejor correlación para el nodo E en el contexto de destino actualmente activo (en este caso,

<PRE_RAÍZ, A', C'>). El proceso luego continúa en el bloque 1306, donde el componente 554 de descodificación selecciona una correlación de transferencia entre el conjunto de correlaciones de transferencia que cubren el nodo E. Por ejemplo, el componente 554 de transferencia selecciona la correlación 1700 de transferencia ilustrada en la FIG. 17. El componente 554 de descodificación avanza luego al bloque 1308 y decide si la correlación de transferencia deja algún nodo filial no cubierto.

Según la LF 1400 de origen, el nodo E no tiene hijos. Por tanto, el componente 554 de descodificación avanza al bloque 1312 para calcular un puntaje para la correlación de transferencia seleccionada. Este puntaje se calcula usando la Ecuación 8, según lo descrito anteriormente, incorporando todos los modelos anteriormente descritos. Obsérvese que un motivo para adoptar el enfoque de arriba hacia abajo de la FIG. 13 es garantizar que el contexto de los nodos (en la correlación que se está puntuando) sea conocido, de modo que el modelo de destino (que requiere el contexto) pueda ser usado para calcular el puntaje del modelo de destino.

Después de puntuar la correlación, el componente 554 de descodificación avanza al bloque 1314 y determina si hay o no alguna correlación más de transferencia que cubra el nodo seleccionado. En este ejemplo, la FIG. 18 ilustra otra correlación 1800 de transferencia para el nodo E seleccionado. Si hay otra correlación de transferencia, el componente 554 de descodificación regresa al bloque 1306 y selecciona la correlación de transferencia adicional. Por ejemplo, la correlación 1800 sería seleccionada. En este ejemplo, la correlación 1800 no tiene ningún nodo hijo no cubierto. Por tanto, el componente 554 de descodificación atraviesa el bloque 1308 hasta el bloque 1312, donde el componente 554 de descodificación calcula el puntaje de la correlación 1800 de transferencia usando la Ecuación 3.

El componente 554 de descodificación avanza luego al bloque 1314 para determinar si hay más correlaciones de transferencia para el nodo seleccionado. En este ejemplo, la FIG. 19 ilustra la correlación 1900 de transferencia que cubre el nodo E. Nuevamente, el componente de descodificación regresa al bloque 1306. En este ejemplo, la correlación 1900 de transferencia no tiene ningún nodo hijo no cubierto. Por tanto, el componente 554 de descodificación calcula un puntaje para la correlación 1900 de transferencia usando la Ecuación 3. Después de que se calcula el puntaje, el componente 554 de descodificación avanza al bloque 1314.

Si el componente 554 de descodificación determina que no hay más correlaciones en la etapa 1314, el proceso continúa en la etapa 1316, donde compara y selecciona la correlación de transferencia con el puntaje más alto, entre las correlaciones de transferencia que cubren el nodo seleccionado. En el ejemplo anterior, se comparan los puntajes para las correlaciones 1700, 1800 y 1900 y se selecciona la correlación con el puntaje más alto. En este ejemplo, se supone que la correlación de transferencia con el puntaje más alto es la correlación 1700 de transferencia. El componente 554 de descodificación almacena el nodo en la cabecera de la correlación, el contexto de la correlación de más alto puntaje (el nodo desde el cual se extiende la correlación en la LF de origen), el puntaje para la correlación de más alto puntaje y la probabilidad de cada modelo individual, o los puntajes de fuentes de información para la correlación de más alto puntaje. De este modo, la probabilidad del modelo de destino, la probabilidad del modelo de canal, la probabilidad del modelo de fertilidad, el puntaje del tamaño y el puntaje de clasificación para la correlación seleccionada son todos almacenados. Aunque se almacena cada probabilidad o puntaje para cada modelo, algún experto en la técnica reconocerá que el puntaje determinado en la Ecuación 8 es el puntaje más importante para almacenar.

Después de que los puntajes para la correlación seleccionada han sido almacenados, el componente 554 de descodificación avanza al bloque 1318, donde determina si existen más niveles para el nodo seleccionado. En el ejemplo anterior, el nodo C está por encima del nodo E. Si hay otro nivel de nodos por encima del nodo actual seleccionado, el componente 554 de descodificación vuelve a la última correlación que estuvo en consideración para ese nodo en el bloque 1320. En el ejemplo anterior, esto implica volver a la correlación 1600 de la FIG. 16.

En el bloque 1322, el componente 554 de descodificación determina si esta correlación tiene algún otro nodo hijo no cubierto que no haya sido explorado. Si hay nodos hijos no cubiertos adicionales para explorar, el componente 554 de descodificación continúa en el bloque 1310, donde el componente 554 de descodificación selecciona el nodo hijo no cubierto. En el ejemplo anterior, esto implicaría seleccionar el nodo hijo F. El componente 554 de descodificación avanza luego al bloque 1304 para determinar si ha sido identificada una correlación óptima para este nodo, dado su contexto. Si no ha sido determinada una correlación óptima, el componente 554 de descodificación selecciona una correlación de transferencia que cubra el nodo hijo seleccionado en la etapa 1306 (p. ej., la correlación 2000 de la FIG. 20). En este ejemplo, la correlación 2000 de transferencia no ha sido puntuada anteriormente. En el bloque 1308, el componente 554 de descodificación determina que el nodo F no tiene ningún nodo hijo no cubierto. Por tanto, el componente 554 de descodificación avanza al bloque 1312 y calcula un puntaje para el nodo F, usando la Ecuación 3.

El componente 554 de descodificación avanza al bloque 1314 para determinar si el nodo F tiene más correlaciones de transferencia. En este ejemplo, ninguna otra correlación de transferencia cubre el nodo F. Por tanto, el componente 554 de descodificación almacena el puntaje para la correlación 2000 de transferencia y almacena cada probabilidad de modelo individual, o el puntaje de origen de la información, para la correlación 2000, los n-a nodos del contexto del sector de destino en el cual fue evaluada la correlación 2000 de transferencia, el nodo de entrada correspondiente al

nodo de cabecera de la correlación 2000 de transferencia y el puntaje total para la correlación 2000 de transferencia.

En este ejemplo, existen más niveles de la LF 1400 de origen por encima del nodo F. Por tanto, el componente 554 de descodificación avanza al bloque 1320 y vuelve a la última correlación para el nodo C que estuvo en consideración. En el bloque 1322, el componente 554 de descodificación determina que la correlación seleccionada para el nodo C no tiene más hijos no cubiertos. Por tanto, el componente 554 de descodificación avanza al bloque 1312 y calcula un puntaje total para la correlación 1600 de transferencia.

Si la correlación seleccionada tuviera hijos no cubiertos, el puntaje para la correlación se determina combinando los puntajes para las correlaciones de más alto puntaje, para los nodos filiales no cubiertos, con el puntaje para la correlación seleccionada. Por ejemplo, los puntajes para la correlación 1700 y 2000 serían combinados con el puntaje para la correlación 1600, para proporcionar un puntaje total para la correlación entera por debajo del nodo C en la LF de origen.

Según una disposición, cada componente de los puntajes de correlaciones se combina por separado. De tal modo, la probabilidad total del modelo de destino de la correlación 1600 es:

Ecuación 17

$$T_{\mu T} = \log P_{\mu T}(m_{1700}) + \log P_{\mu T}(m_{2000}) + \log P_{\mu T}(m_{1600})$$

donde $P_{\mu T}(m_{1700})$ es la probabilidad del modelo de destino para la correlación 1700, $P_{\mu T}(m_{2000})$ es la probabilidad del modelo de destino para la correlación 2000 y $\log P_{\mu T}(m_{1600})$ es la probabilidad del modelo de destino para la correlación 1600.

De manera similar, la probabilidad total de modelo de canal de la correlación 1600 es:

Ecuación 18

$$T_{\mu C} = \log P_{\mu C}(m_{1700}) + \log P_{\mu C}(m_{2000}) + \log P_{\mu C}(m_{1600})$$

y la probabilidad total del modelo de fertilidad de la correlación 1600 es:

Ecuación 19

$$T_{\mu F} = \log P_{\mu F}(m_{1700}) + \log P_{\mu F}(m_{2000}) + \log P_{\mu F}(m_{1600})$$

El puntaje total del tamaño de correlación para la correlación 1600 es el promedio de los puntajes filiales del tamaño de correlación y el puntaje del tamaño de correlación para la correlación 1600 solamente, de modo que:

Ecuación 20

$$S_{\mu S} = [Puntaje_{\mu S}(m_{1700}) + Puntaje_{\mu S}(m_{2000}) + Puntaje_{\mu S}(m_{1600})] / 3$$

Como el puntaje total del tamaño de correlación, el puntaje total de clasificación de la correlación 1600 es el promedio de los puntajes filiales de clasificación y el puntaje de clasificación para la correlación 1600 solamente, y se describe como la:

Ecuación 21

$$S_{\mu B} = [Puntaje_{\mu B}(m_{1700}) + Puntaje_{\mu B}(m_{2000}) + Puntaje_{\mu B}(m_{1600})] / 3$$

Una vez que han sido determinados los puntajes totales para cada componente, se combinan en un único puntaje para la correlación seleccionada, usando la Ecuación 8 anterior.

El componente 554 de descodificación avanza luego al bloque 1314 y decide si existen o no más correlaciones de transferencia para el nodo C. En este ejemplo, no existe ninguna otra correlación de transferencia para el nodo C, por lo que el componente 554 de descodificación selecciona la correlación 1600 de transferencia como la que tiene la más alta correlación de puntaje, y almacena el puntaje total para la correlación 1600, el contexto (nodo A) de la correlación 1600, el nodo de cabecera de la correlación 1600 (nodo C) y los puntajes componentes individuales totales para la correlación 1600.

En el bloque 1318, el componente 554 de descodificación decide si existen o no más niveles por encima del nodo C en la LF 1400 de origen. En este ejemplo, el nodo A está por encima del nodo C. Por tanto, el componente 554 de descodificación vuelve al próximo nivel superior en las correlaciones, según lo ilustrado en el bloque 1320. En el ejemplo anterior, esto implica volver a la correlación 1500 de la FIG. 15. En el bloque 1322, el componente 554 de descodificación determina si la correlación seleccionada tiene algún otro nodo hijo no cubierto que necesite ser

explorado. En este ejemplo, no hay ningún otro nodo hijo no cubierto, por lo que el componente 554 de descodificación avanza al bloque 1312 y calcula un puntaje total para la correlación 1500 de transferencia. Como la correlación 1600, el puntaje total para la correlación 1500 de transferencia está formado por la combinación de los puntajes para la correlación 1600 con los puntajes para la correlación 1500.

El componente 554 de descodificación avanza entonces al bloque 1314 y determina si existen más correlaciones de transferencia para el nodo A. En este ejemplo, la correlación 2100 de transferencia también cubre el nodo A. Como resultado, el proceso vuelve a la etapa 1306 para seleccionar la correlación 2100 de transferencia.

En la etapa 1308, el proceso determina que la correlación 2100 tiene un nodo hijo no cubierto. Específicamente, los nodos E y F no están cubiertos por la correlación 2100. En la etapa 1310, el nodo E es seleccionado y el proceso vuelve a la etapa 1304 para determinar si ha sido seleccionada una correlación óptima para el nodo E en el contexto actual, dada nuestra elección anterior de la correlación 2100, que en este caso sería <PRE_RAÍZ, A', C'>. Una correlación óptima de ese tipo fue seleccionada (correlación 1700). Esta correlación óptima y sus puntajes son luego seleccionados y el proceso vuelve a la correlación 2100 en la etapa 1320.

El proceso determina luego si hay más nodos filiales no cubiertos para considerar. Para la correlación 2100, el nodo hijo F no ha sido considerado y es seleccionado en la etapa 1310. En la etapa 1304, se determina que ha sido determinada una correlación óptima para el nodo F en el contexto del nodo D (correlación 2000). Esta correlación óptima es seleccionada luego y el proceso vuelve a la correlación 2100 en la etapa 1320.

Al volver a la etapa 1322, no hay más nodos filiales no cubiertos a considerar, y se calcula un puntaje para la correlación 2100, usando los puntajes almacenados para las correlaciones 1700 y 2000 y los puntajes de correlaciones individuales para la correlación 2100. Como antes, los componentes individuales del puntaje de la correlación son combinados por separado.

En la etapa 1314, no existe ninguna otra correlación de transferencia, por lo que el componente 554 de descodificación avanza al bloque 1316 y selecciona entre la estructura de correlación de transferencia encabezada por la correlación 1500 de transferencia y la estructura de correlación de transferencia encabezada por la correlación 2100 de transferencia, en base a los puntajes totales para estas dos estructuras de correlación. El componente 554 de descodificación almacena el puntaje total para la estructura de correlación de transferencia con el puntaje más alto y avanza al bloque 1318. En el bloque 1318, el componente de descodificación determina si existen o no más niveles por encima del nodo A. En este ejemplo, el nodo A es el nodo supremo en la LF 1400 de origen, por lo que el componente 554 de descodificación termina la descodificación y devuelve la estructura de correlación de transferencia de más alto puntaje determinada para el nodo A.

En base a la correlación de transferencia almacenada con el mayor puntaje, ilustrada en las FIGS. 15 a 21, el componente 526 de transferencia puede construir una LF de destino. Por ejemplo, si las correlaciones 1500, 1600, 1700 y 1900 de transferencia fueron seleccionadas como las correlaciones de transferencia con más alto puntaje y, por lo tanto, tienen la más alta probabilidad de una traducción de destino, son combinadas para formar una LF de destino.

Las fuentes de información, tales como los modelos estadísticos y otras técnicas de puntaje, pueden ser usadas para determinar la mejor traducción para una estructura semántica. Las estructuras semánticas de entrada han sido usadas para generar estructuras semánticas de salida, usando un algoritmo de búsqueda voraz sobre un conjunto de correlaciones de transferencia. Sin embargo, el algoritmo de búsqueda voraz no prueba todas las combinaciones posibles de las correlaciones de transferencia, sino que sencillamente selecciona el primer conjunto de correlaciones de transferencia que cubre completamente la estructura semántica de entrada. Modelos estadísticos han sido usados para predecir la cadena de salida más probable, dada una cadena de entrada. Sin embargo, los modelos estadísticos usados en los sistemas basados en cadenas suponen que un elemento puede ser predicho en base a otro elemento adyacente, o casi adyacente. De tal modo, la presente invención usa modelos estadísticos para predecir la mejor traducción para una estructura semántica.

Aunque el modelo de idioma de salida, según se aplica a estructuras semánticas de la presente invención, puede ser usado en la construcción de cadenas de palabras de salida, el modelo de idioma de destino, según se aplica a estructuras semánticas, puede ser usado en otros programas de idiomas. Por ejemplo, otros sistemas incluyen el reconocimiento del habla, el reconocimiento de caracteres ópticos, el reconocimiento de la escritura manual, la extracción de información y la corrección gramatical.

Aunque la presente invención ha sido descrita con referencia a realizaciones específicas, los expertos en la técnica reconocerán que pueden hacerse cambios en la forma y el detalle sin apartarse del alcance de la invención.

REIVINDICACIONES

1. Un procedimiento implementado por ordenador de descodificación durante un tiempo de ejecución de una traducción, de una estructura (552) semántica de entrada representada por una forma lógica (552) de origen obtenida de un texto (550) de origen en un primer idioma, para generar una estructura (556) semántica de salida representada por una forma lógica (556) de salida, comprendiendo el procedimiento:

hallar un primer conjunto de correlaciones (800; 900; 1000; 1200; 1500; 1600; 1700; 1800; 1900; 2000; 2100) de transferencia contenidas en una base (518) de datos de correlaciones de transferencia, que fue formada durante el entrenamiento con un corpus bilingüe, teniendo cada correlación de transferencia un sector semántico de entrada que describe nodos de la estructura semántica de entrada, y teniendo un sector semántico de salida que describe nodos de la estructura semántica de salida;

calcular (1312) un puntaje para las correlaciones de transferencia en el conjunto de correlaciones de transferencia que cubren un nodo seleccionado de la estructura semántica de entrada usando un modelo estadístico,

en el cual el cálculo de un puntaje para las correlaciones de transferencia comprende

seleccionar (1306), entre el conjunto de correlaciones de transferencia, una correlación de transferencia que tenga un sector de origen que cubra el nodo seleccionado;

determinar (1308) si hay algún nodo hijo en la estructura semántica de entrada que no esté cubierto por la correlación seleccionada y que se extienda directamente desde la correlación seleccionada,

y, si es así, combinar puntajes para las correlaciones de más alto puntaje, para los nodos filiales no cubiertos, con el puntaje para la correlación seleccionada;

seleccionar (1316) la correlación de transferencia con el más alto puntaje, en base al puntaje;

determinar (1318) si existen más niveles por encima del nodo seleccionado, a fin de determinar (1322) si la correlación seleccionada tiene algún otro nodo hijo no cubierto y, si es así, seleccionar (1310) este nodo hijo no cubierto como un nuevo nodo seleccionado, y repetir las etapas del procedimiento, a partir de la etapa del hallazgo; y si no existen más niveles por encima del nodo seleccionado, se seleccionan las correlaciones de transferencia con el más alto puntaje

usar las correlaciones de transferencia seleccionadas para construir la estructura semántica de salida; y

generar un texto (558) de destino en un segundo idioma, en base a la estructura (556) semántica de salida.

2. El procedimiento de la reivindicación 1, en el cual el cálculo de un puntaje para al menos una correlación de transferencia comprende calcular un puntaje usando un modelo de idioma de destino que proporcione una probabilidad de que un conjunto de nodos aparezca en la estructura semántica de salida.

3. El procedimiento de la reivindicación 1, en el cual el cálculo de un puntaje para al menos una correlación de transferencia comprende calcular un puntaje usando un modelo de canal que proporcione una probabilidad de un sector semántico de entrada de una correlación de transferencia, dado el sector semántico de salida de la correlación de transferencia.

4. El procedimiento de la reivindicación 3, en el cual el cálculo de un puntaje usando el modelo de canal comprende normalizar un puntaje del modelo de canal en base a un cierto número de nodos solapados entre las correlaciones de transferencia.

5. El procedimiento de la reivindicación 1, en el cual el cálculo de un puntaje para el menos una correlación de transferencia comprende calcular un puntaje usando un modelo de fertilidad que proporcione una probabilidad de borrar un nodo en una correlación de transferencia.

6. El procedimiento de la reivindicación 1, en el cual el cálculo de un puntaje para al menos una correlación de transferencia comprende calcular un puntaje de tamaño en base a un cierto número de nodos en el sector semántico de entrada de la correlación de transferencia.

7. El procedimiento de la reivindicación 1, en el cual el cálculo de un puntaje para al menos una correlación de transferencia comprende calcular un puntaje de clasificación en base a un cierto número de características binarias coincidentes en la estructura semántica de entrada y el sector semántico de entrada de la correlación de transferencia.

8. El procedimiento de la reivindicación 1, en el cual el cálculo de un puntaje para al menos una correlación de transferencia en el conjunto de correlaciones de transferencia comprende:

calcular puntajes distintos para una pluralidad de modelos; y

combinar los distintos puntajes para determinar el puntaje para la correlación de transferencia.

9. El procedimiento de la reivindicación 8, en el cual la pluralidad de modelos comprende un modelo de canal que proporciona una probabilidad de un sector semántico de entrada de una correlación de transferencia, dado el sector semántico de salida de la correlación de transferencia.

10. El procedimiento de la reivindicación 8, en el cual la pluralidad de modelos comprende un modelo de fertilidad que proporciona una probabilidad de borrar un nodo en una correlación de transferencia.

11. El procedimiento de la reivindicación 8, en el cual la pluralidad de modelos comprende un modelo de idioma de destino que proporciona una probabilidad de que un conjunto de nodos aparezca en la estructura semántica de salida.

12. El procedimiento de la reivindicación 8, y que comprende adicionalmente:

calcular un puntaje de tamaño para la correlación de transferencia, estando el puntaje de tamaño basado en un cierto número de nodos en el sector semántico de entrada de la correlación de transferencia; y

combinar el puntaje de tamaño con los distintos puntajes para la pluralidad de modelos, para determinar el puntaje para la correlación de transferencia.

13. El procedimiento de la reivindicación 8, y que comprende adicionalmente:

calcular un puntaje de clasificación para la correlación de transferencia, estando el puntaje de clasificación basado en un cierto número de características binarias coincidentes en la estructura semántica de entrada y el sector semántico de entrada de la correlación de transferencia; y

combinar el puntaje de clasificación con los distintos puntajes para la pluralidad de modelos, para determinar el puntaje para la correlación de transferencia.

14. El procedimiento de la reivindicación 8, en el cual la combinación de los puntajes comprende:

multiplicar cada puntaje por un peso, para formar puntajes de modelos ponderados; y

sumar los puntajes de modelos ponderados para determinar el puntaje para cada correlación de transferencia.

15. El procedimiento de la reivindicación 1, en el cual la provisión de un conjunto de correlaciones de transferencia comprende proveer un conjunto de correlaciones de transferencia dispuestas como una estructura de árbol y múltiples niveles de sub-árboles anidados, que comprenden una correlación de transferencia de raíz y sub-árboles, comprendiendo cada sub-árbol una correlación de transferencia de raíz, en donde cada correlación de transferencia en el conjunto de correlaciones de transferencia aparece como una correlación de transferencia de raíz en al menos uno entre el árbol y los sub-árboles.

16. Un sistema (500) de traducción automática para traducir una entrada (550) en un primer idioma, a una salida (558) en un segundo idioma, comprendiendo el sistema:

un analizador sintáctico (522) para analizar sintácticamente la entrada en una representación (552) semántica de entrada, como una forma lógica (534) de origen;

un componente (524) de búsqueda configurado para hallar un conjunto de correlaciones de transferencia, de formas lógicas (534) de origen a formas lógicas (536) de destino, en donde cada correlación de transferencia corresponde a una parte de la representación semántica de entrada, y fue formada durante el entrenamiento con un corpus bilingüe;

un componente (554) de descodificación configurado para puntuar (1312) una pluralidad de correlaciones de transferencia que cubren un nodo seleccionado de la representación semántica de entrada, y para seleccionar (1316) la correlación de transferencia de más alto puntaje, en base a los puntajes,

en el cual la puntuación de la pluralidad de correlaciones de transferencia comprende

seleccionar (1306), entre una pluralidad de correlaciones de transferencia, una correlación de transferencia que tenga un sector de origen que cubra el nodo seleccionado;

determinar (1308) si hay algún nodo hijo en la representación semántica de entrada que no esté cubierto por la correlación seleccionada y que se extienda directamente desde la correlación seleccionada,

y, si es así, combinar los puntajes para las correlaciones de más alto puntaje, para los nodos filiales no cubiertos, con

el puntaje para la correlación seleccionada;

en el cual el componente de descodificación está adicionalmente configurado para determinar (1318) si existen más niveles por encima del nodo seleccionado, a fin de determinar (1322) si la correlación seleccionada tiene algún otro nodo hijo no cubierto y, si es así, seleccionar (1310) este nodo hijo no cubierto como un nuevo nodo seleccionado para el componente de descodificación y, si no existen más niveles por encima del nodo seleccionado, seleccionar las correlaciones de transferencia con los más altos puntajes

y un componente (528) de generación configurado para generar la salida en base a las correlaciones de transferencia seleccionadas con más altos puntajes.

17. El sistema de traducción automática de la reivindicación 16, en el cual el componente de descodificación puntúa cada correlación de transferencia usando una pluralidad de modelos estadísticos.

18. El sistema de traducción automática de la reivindicación 17, en el cual la salida comprende una representación (556) semántica de salida y en el cual la pluralidad de modelos estadísticos comprende un modelo de destino que proporciona una probabilidad de que una secuencia de nodos aparezca en la representación semántica de salida.

19. El sistema de traducción automática de la reivindicación 17, en el cual la pluralidad de modelos estadísticos comprende un modelo de canal que proporciona una probabilidad de un conjunto de nodos semánticos en un sector de entrada de una correlación de transferencia, dado un conjunto de nodos semánticos en un sector de salida de la transferencia.

20. El sistema de traducción automática de la reivindicación 17, en el cual la pluralidad de modelos estadísticos comprende un modelo de fertilidad que proporciona una probabilidad de borrar un nodo en la correlación de transferencia.

21. El sistema de traducción automática de la reivindicación 17, en el cual el componente de descodificación puntúa cada correlación de transferencia usando un puntaje de tamaño, basado en un cierto número de nodos en un sector de entrada de la correlación de transferencia.

22. El sistema de traducción automática de la reivindicación 17, en el cual el componente de descodificación puntúa cada correlación de transferencia usando un puntaje de clasificación basado en un cierto número de características binarias coincidentes entre la entrada y un sector de salida de la correlación de transferencia.

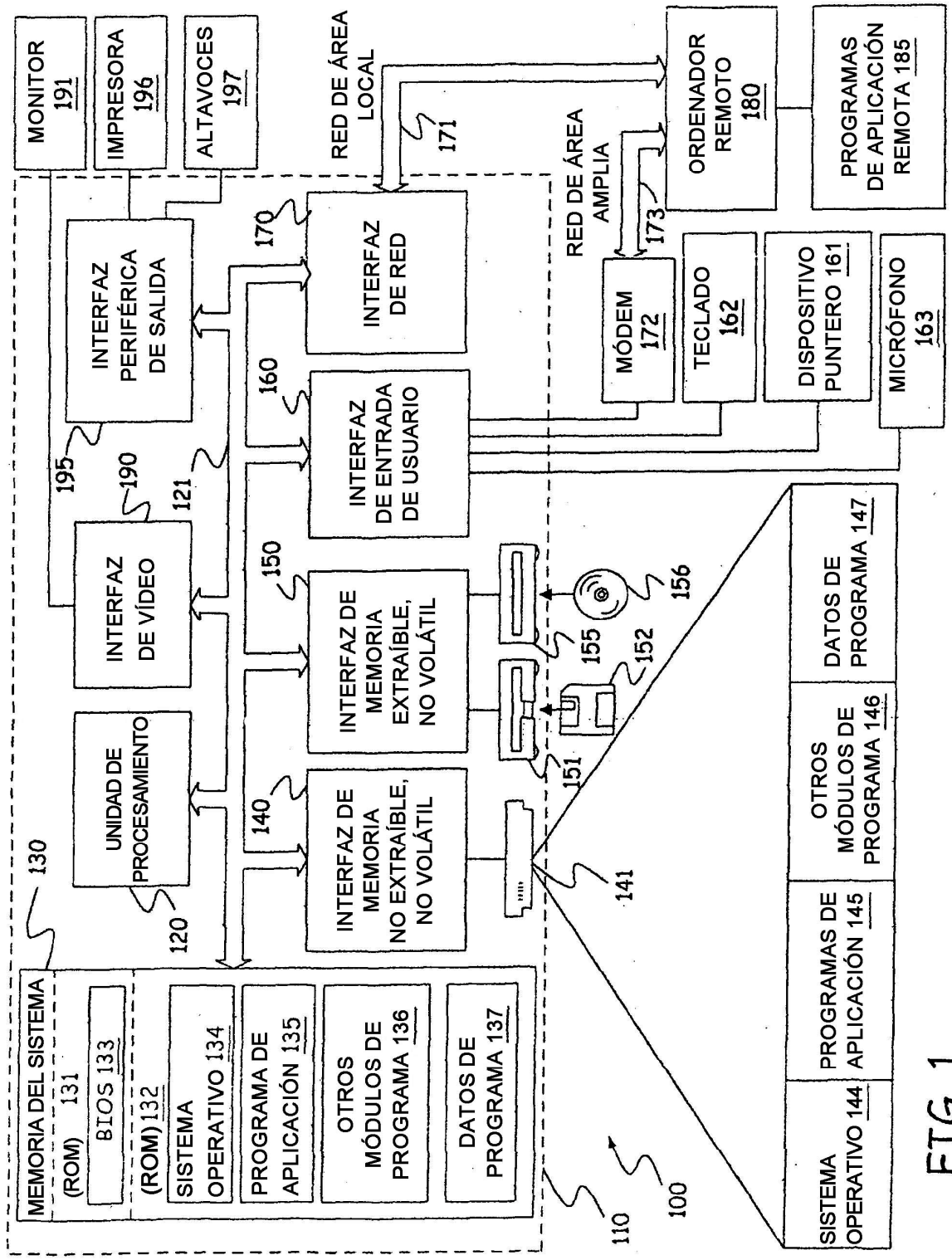

MONITOR 191
IMPRESORA 196
ALTAVOCES 197
RED DE ÁREA LOCAL
171
ORDENADOR REMOTO 180
PROGRAMAS DE APLICACIÓN REMOTA 185
RED DE ÁREA AMPLIA
173
INTERFAZ PERIFÉRICA DE SALIDA
170
INTERFAZ DE RED
MÓDEM 172
TECLADO 162
DISPOSITIVO PUNTERO 161
MICRÓFONO 163
160
INTERFAZ DE ENTRADA DE USUARIO
195
190
121
INTERFAZ DE VÍDEO
150
INTERFAZ DE MEMORIA EXTRAÍBLE, NO VOLÁTIL
156
155
152
151
UNIDAD DE PROCESAMIENTO
140
INTERFAZ DE MEMORIA NO EXTRAÍBLE, NO VOLÁTIL
141
130
120
MEMORIA DEL SISTEMA
(ROM) 131
BIOS 133
(ROM) 132
SISTEMA OPERATIVO 134
PROGRAMA DE APLICACIÓN 135
OTROS MÓDULOS DE PROGRAMA 136
DATOS DE PROGRAMA 137
DATOS DE PROGRAMA 147
OTROS MÓDULOS DE PROGRAMA 146
PROGRAMAS DE APLICACIÓN 145
SISTEMA OPERATIVO 144
110
100

FIG. 1

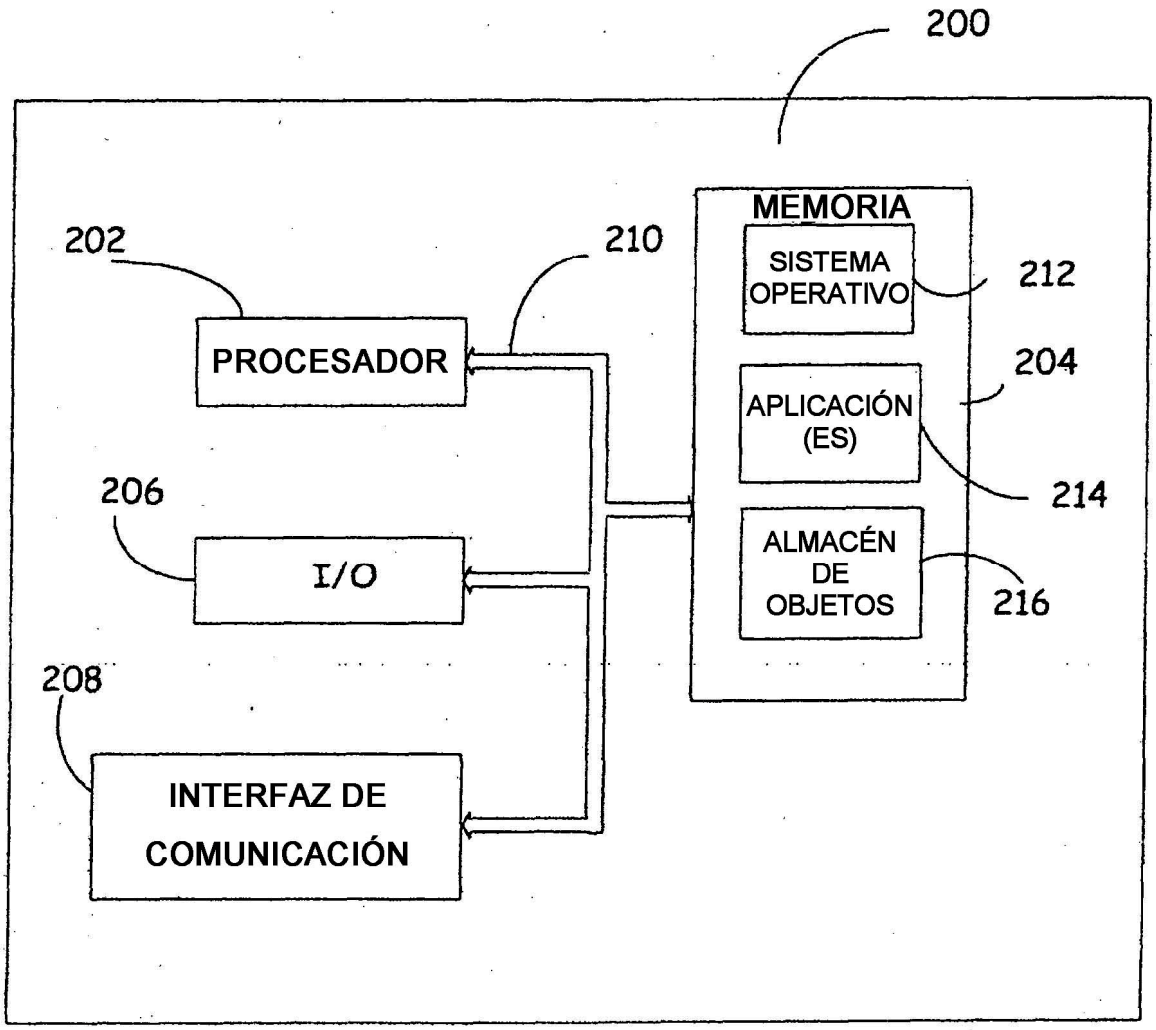

200
MEMORIA
SISTEMA OPERATIVO
212
APLICACIÓN (ES)
204
214
ALMACÉN DE OBJETOS
216
202
PROCESADOR
210
206
I/O
208
INTERFAZ DE COMUNICACIÓN

FIG. 2

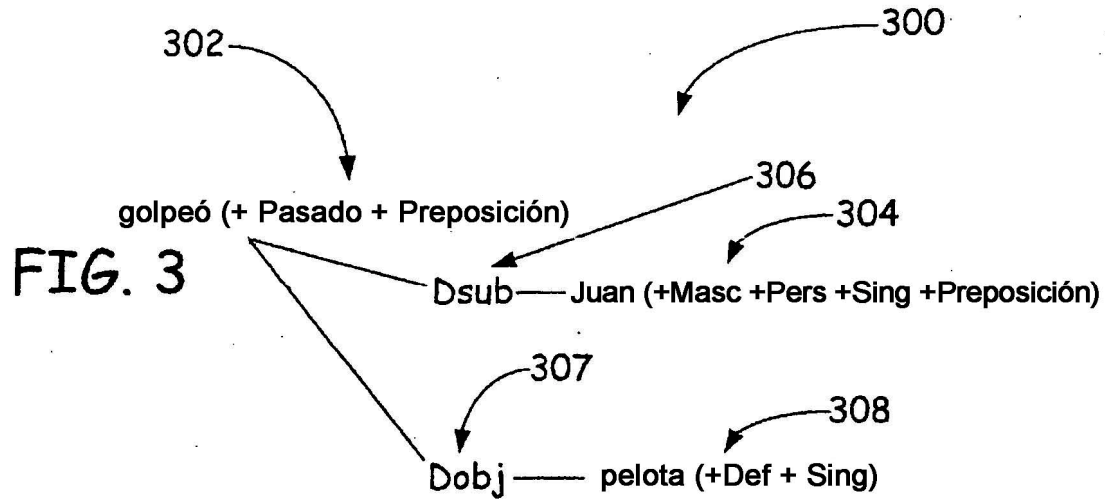

302
300
306
304
golpeó (+ Pasado + Preposición)
Dsub
Juan (+Masc +Pers +Sing +Preposición)
307
308
Dobj
pelota (+Def + Sing)

FIG. 3

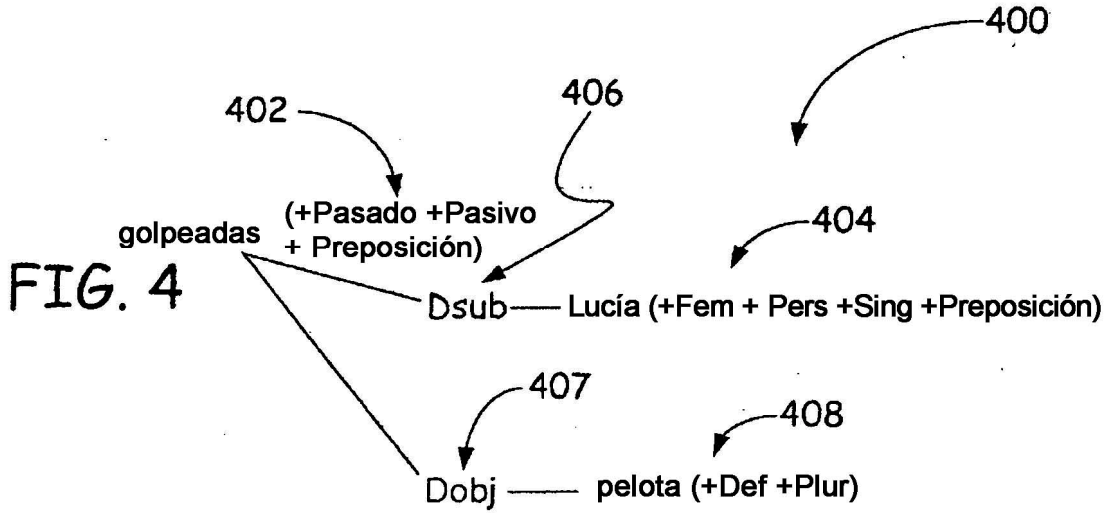
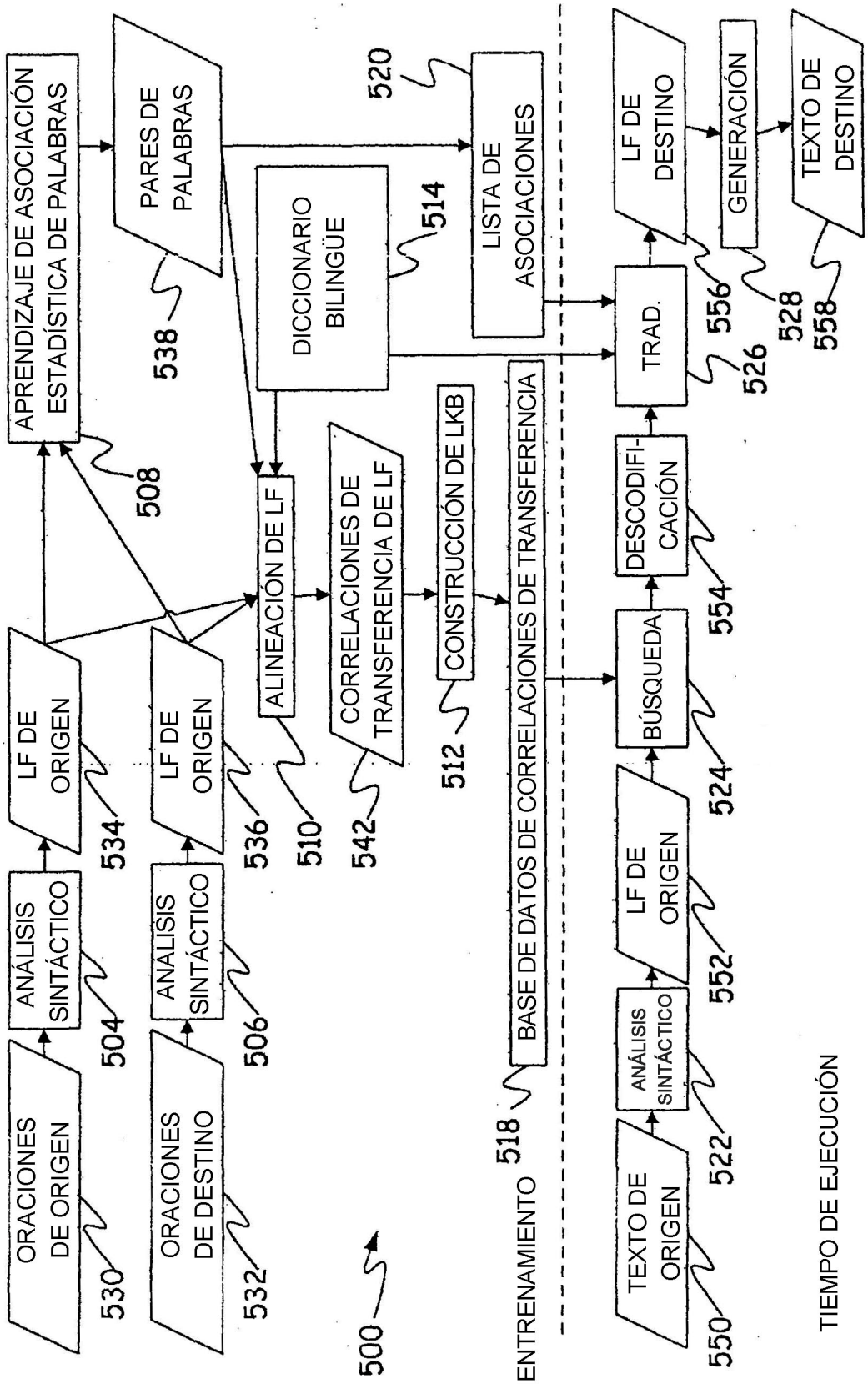

400
402
406
(+Pasado +Pasivo
+ Preposición)
golpeadas
404
Dsub
Lucía (+Fem + Pers +Sing +Preposición)
407
408
Dobj
pelota (+Def +Plur)

FIG. 4

FIG. 5


ORACIONES DE ORIGEN
530
ANÁLISIS SINTÁCTICO
504
LF DE ORIGEN
534
ORACIONES DE DESTINO
532
ANÁLISIS SINTÁCTICO
506
LF DE ORIGEN
536
APRENDIZAJE DE ASOCIACIÓN ESTADÍSTICA DE PALABRAS
508
PARES DE PALABRAS
538
DICCIONARIO BILINGÜE
514
520
LISTA DE ASOCIACIONES
ALINEACIÓN DE LF
510
CORRELACIONES DE TRANSFERENCIA DE LF
542
CONSTRUCCIÓN DE LKB
512
BASE DE DATOS DE CORRELACIONES DE TRANSFERENCIA
518
500
ENTRENAMIENTO
TEXTO DE ORIGEN
550
ANÁLISIS SINTÁCTICO
522
LF DE ORIGEN
552
BÚSQUEDA
524
DESCODIFI-CACIÓN
554
TRAD.
526
LF DE DESTINO
556
GENERACIÓN
528
TEXTO DE DESTINO
558
TIEMPO DE EJECUCIÓN

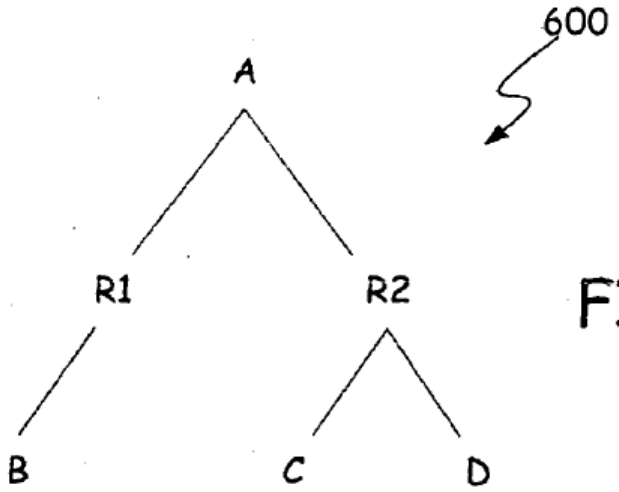


FIG. 6

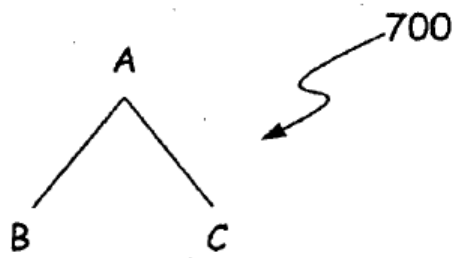


FIG. 7

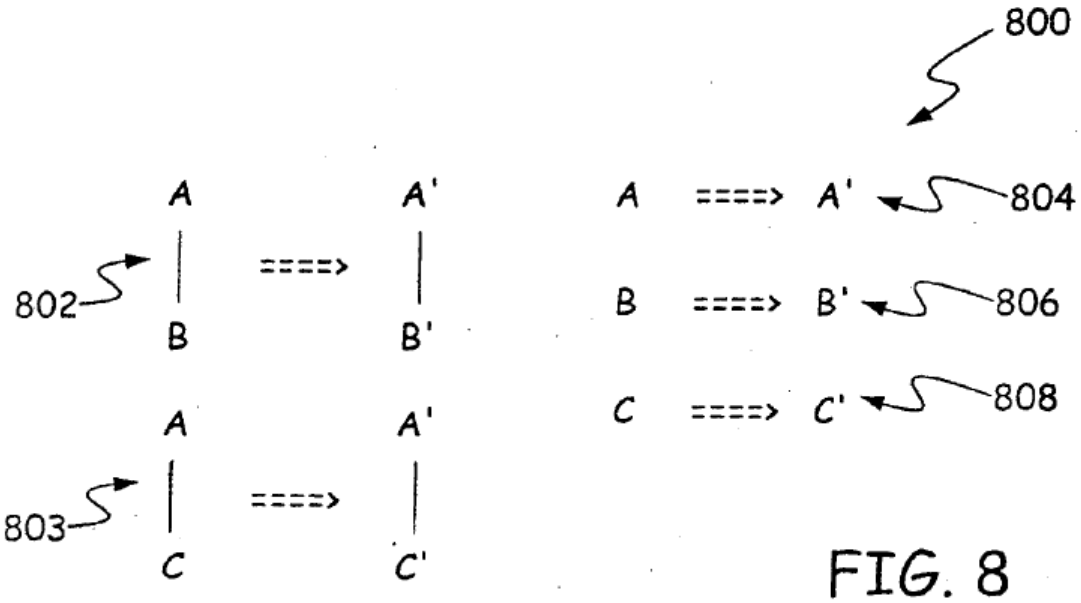


FIG. 8

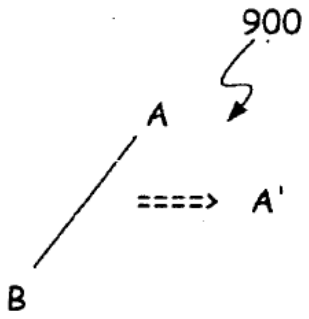


FIG. 9

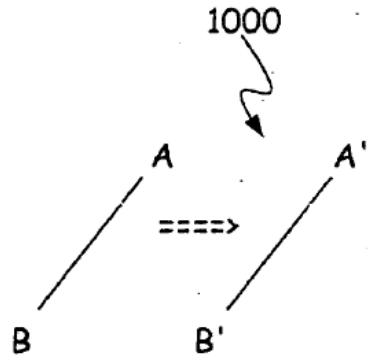


FIG. 10

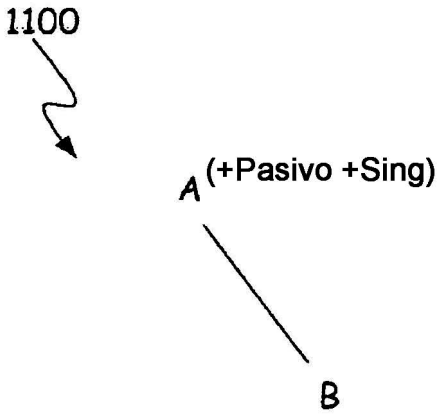


FIG. 11

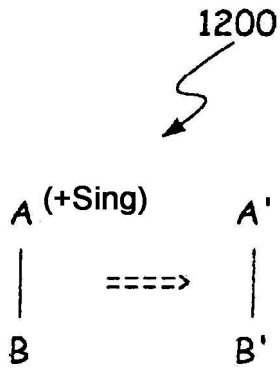


FIG. 12

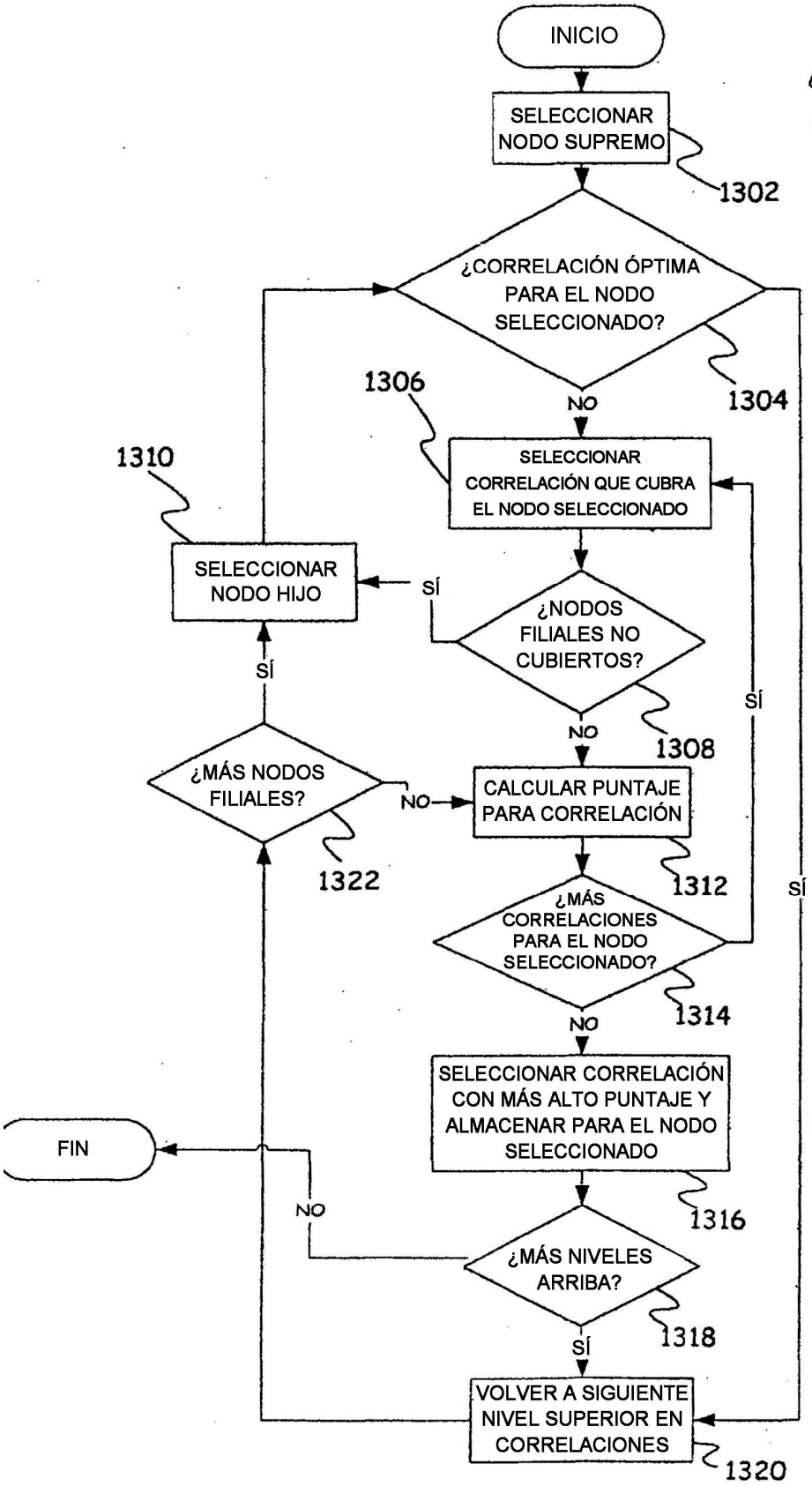

INICIO
1300
SELECCIONAR NODO SUPREMO
1302
¿CORRELACIÓN ÓPTIMA PARA EL NODO SELECCIONADO?
1304
NO
SÍ
1306
SELECCIONAR CORRELACIÓN QUE CUBRA EL NODO SELECCIONADO
¿NODOS FILIALES NO CUBIERTOS?
SÍ
NO
1308
1310
SELECCIONAR NODO HIJO
SÍ
¿MÁS NODOS FILIALES?
NO
1322
CALCULAR PUNTAJE PARA CORRELACIÓN
1312
¿MÁS CORRELACIONES PARA EL NODO SELECCIONADO?
SÍ
NO
1314
SELECCIONAR CORRELACIÓN CON MÁS ALTO PUNTAJE Y ALMACENAR PARA EL NODO SELECCIONADO
1316
FIN
NO
¿MÁS NIVELES ARRIBA?
1318
SÍ
VOLVER A SIGUIENTE NIVEL SUPERIOR EN CORRELACIONES
1320

FIG. 13

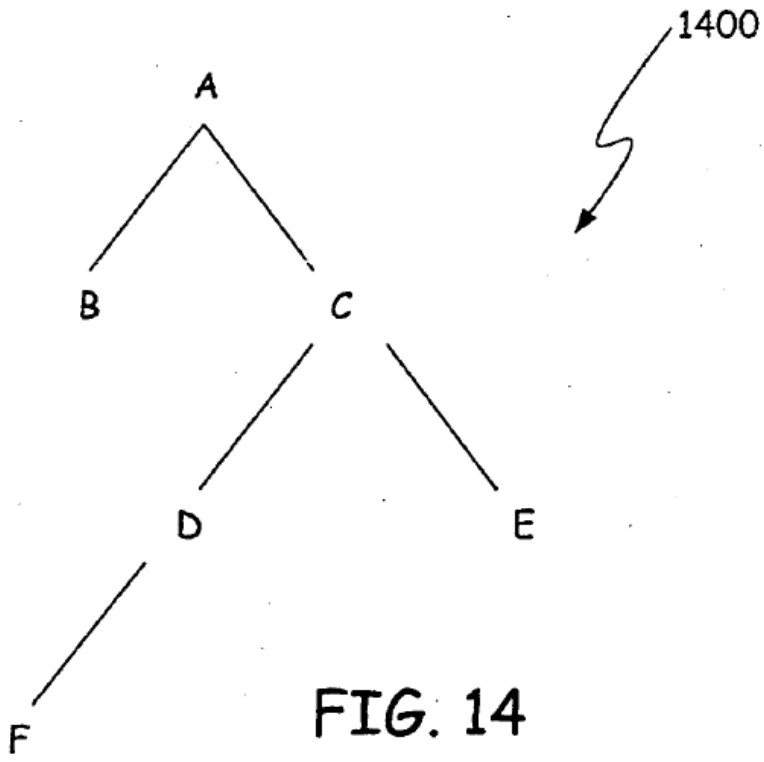


FIG. 14

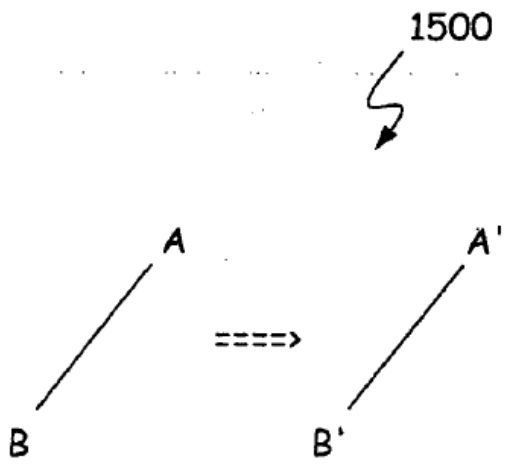


FIG. 15

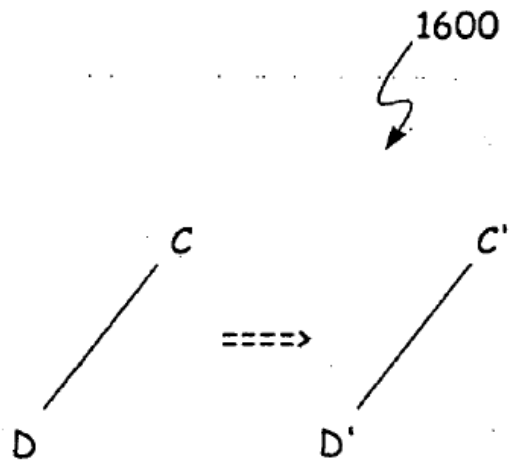


FIG. 16

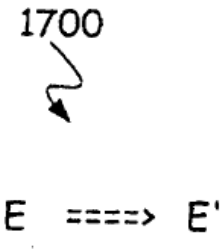


FIG. 17

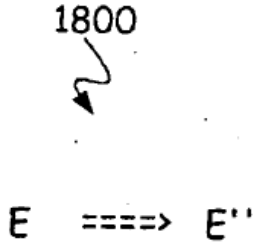


FIG. 18

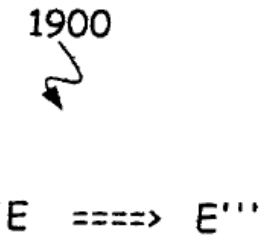


FIG. 19

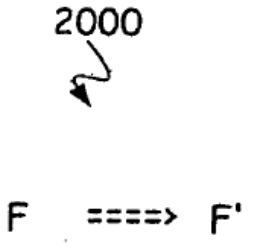


FIG. 20

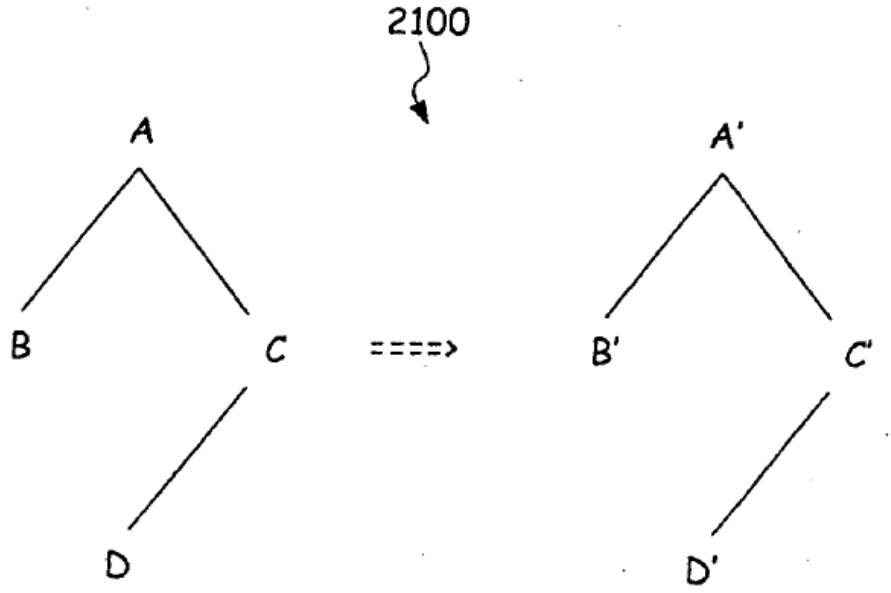


FIG. 21