Alineación de textos y traducción automática

ITALICA
Universidad de Sevilla
José A. Troyano

Índice

- **Introducción**
- Alineamiento sin usar información léxica
- Alineamiento usando información léxica
- Traducción automática

Introducción

Definiciones

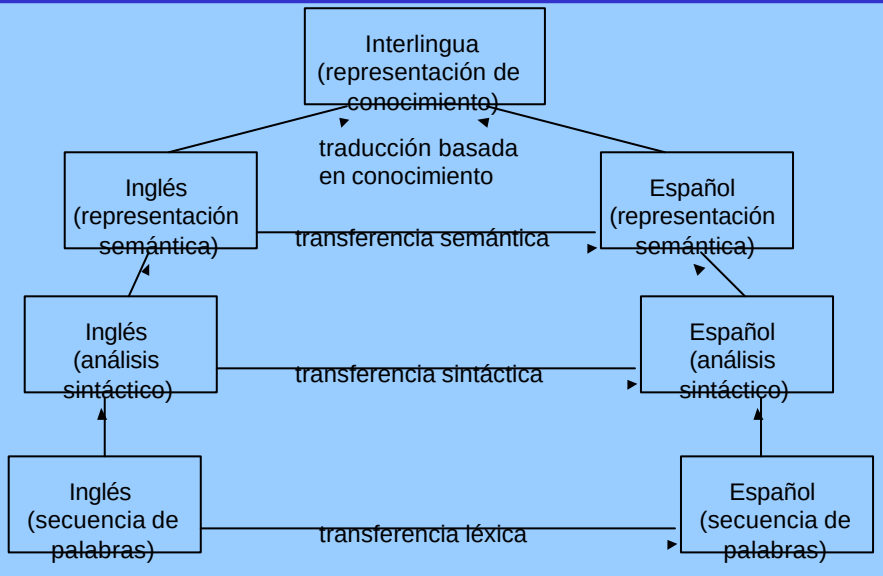
Traducción automática: Obtener un texto correcto y fluido equivalente a un texto en otro lenguaje.

Las herramientas actuales distan mucho de eso (excepto para dominios muy restringidos como por ejemplo los informes meteorológicos).

Alineamiento: Dados dos textos (bitextos) habitualmente en idiomas distintos, identificar qué partes de uno (párrafos, frases, palabras) corresponden con las del otro.

Introducción

Estrategias en la traducción automática (I)



Introducción

Estrategias en la traducción automática (II)

Transferencia léxica: problemas con el orden de las palabras

Where are you from? ? Donde eres tú de?

(*De donde eres tú?*)

Transferencia sintáctica: problemas con las construcciones gramaticales

en mala hora ? at bad time

(*unluckily*)

Transferencia semántica: problemas con las funciones de los sintagmas

La botella entró ? The bottle entered the cave floating

en la cueva flotando (*the bottle floated into the cave*)

Introducción

Alineación de frases

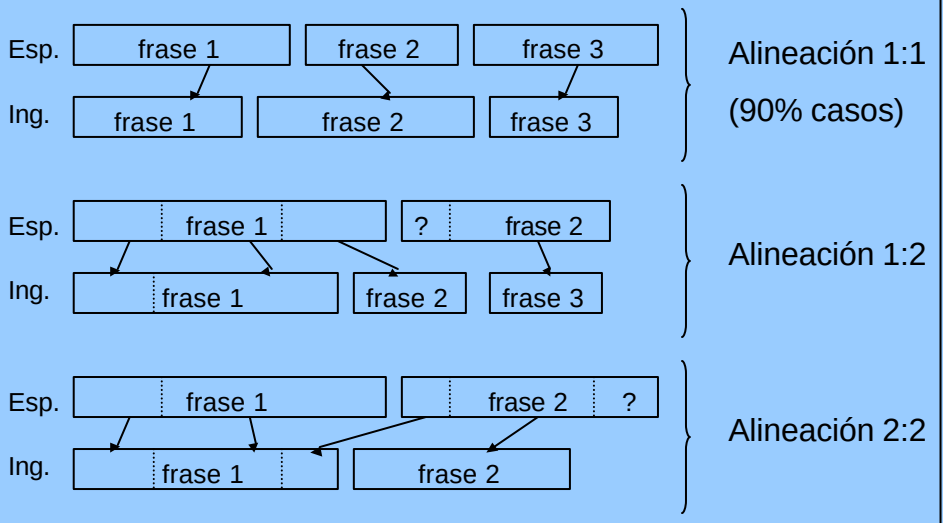
Se trata de buscar beads (cuentas de collar). Un bead es el emparejamiento de un grupo de frases en el corpus del lenguaje origen con otro grupo de frases en el corpus del lenguaje destino.

Es una forma de “etiquetar” un corpus bilingüe. Un bitexto alineado es un recurso lingüístico muy útil en las tareas relacionadas con la traducción automática.

Se suelen utilizar actas parlamentarias de países o estados multiligües, como Canadá, Hong-Kong, Unión Europea,...

Introducción

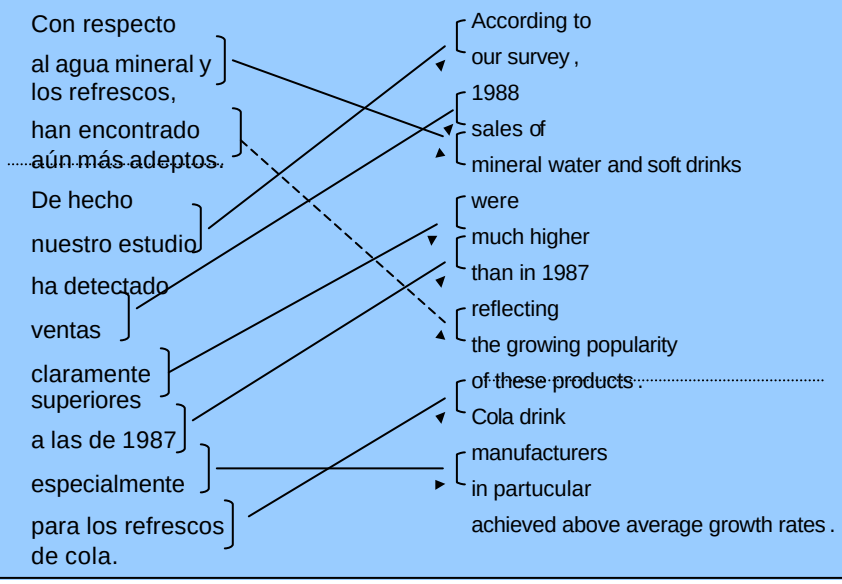
Posibles alineaciones de frases



Lo normal es encontrar 0:1, 1:0, 1:1, 1:2, 2:1, 2:2, 2:3, 3:2

Introducción

Ejemplo de alineación 2:2



Introducción

Alineamiento y correspondencia

Además de los problemas de alineamiento también puede haber problemas de cambio de orden de las frases:

• Alineamiento: Sólo resuelve la identificación de fragmentos. Se asume que el orden de las frases se respeta.

• Correspondencia: Problema más complejo (alineación + cambio de orden).

Los métodos que vamos a ver sólo abordan el primer problema.

Índice

Alineamiento sin usar información léxica

Métodos basados en longitud

En estos métodos se desecha la información léxica y la única información utilizada es la longitud de las unidades de texto.

Se busca el mejor alineamiento A, dados los textos S y T:

$$\arg_A \max P(A|S,T) = \arg_A \max P(A,S,T)$$

Un alineamiento es una secuencia de cuentas $(B_1,...,B_k)$, asumiendo que son independientes nos queda:

$$P(A,S,T) \; ? \; ? \; _{k=1..K} \; P(B_k)$$

La cuestión es calcular $P(B_k)$ dadas las frases de los textos.

Alineamiento sin usar información léxica

Métodos basados en longitud: Gale-Church (I)

Este método calcula $P(B_k)$ en base a la longitud (en caracteres) de las frases. El método requiere que los textos estén alineados a nivel de párrafos.

Sólo contempla los alineamientos siguientes:

1:1, 1:0, 0:1, 2:1, 1:2, 2:2

Dados los textos $s_1,...,s_i$ y $t_1,...,t_j$, se define el coste de su alineamiento con la siguiente distancia:

$$D(i,j)$$

Calculando el mínimo de esta distancia obtendremos el mejor alineamiento.

Alineamiento sin usar información léxica

Métodos basados en longitud: Gale-Church (II)

$$D(i,j) = \min \begin{cases} D(i,j-1) + \text{coste}(0{:}1, |?|, |t_j|) \\ D(i-1,j) + \text{coste}(1{:}0, |s_i|, |?|) \\ D(i-1,j-1) + \text{coste}(1{:}1, |s_i|, |t_j|) \\ D(i-2,j-1) + \text{coste}(2{:}1, |s_{i-1}, s_i|, |t_j|) \\ D(i-1,j-2) + \text{coste}(2{:}2, |s_i|, |t_{j-1}, t_j|) \\ D(i-2,j-2) + \text{coste}(1{:}2, |s_{i-1}, s_i|, |t_{j-1}, t_j|) \end{cases}$$

Se asume que cada carácter en un lenguaje produce de forma aleatoria un número de caracteres en otro lenguaje (distribución normal con parámetros ? y s^2 estimados sobre el corpus).

El coste se calcula así:

coste(tipo_alin,long1,long2) = -log P(tipo_alin| ?(long1, long2))

?(l1, l2) = (l2 - l1?)/raiz(l1s^2)

La probabilidad P(tipo_alin |?) se calcula mediante la regla de Bayes en base a los datos del corpus de entrenamiento.

Alineamiento sin usar información léxica

Métodos basados en longitud: Brown y Wu

Brown: Una adaptación del método de Gale y Church midiendo la distancia en palabras en lugar de en caracteres.

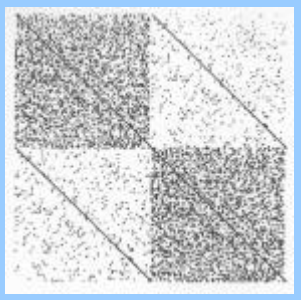
Wu: Demuestra que la asunciones de Gale y Church no son válidas para lenguajes dispares (p.e. inglés y chino). Amplía dicho método con pistas léxicas.

Alineamiento sin usar información léxica

Alineamiento basado en *cognados*: Church

Cognado: palabra con similar estructura en dos lenguajes.

El método busca pares de secuencias de caracteres (p.e. tetragramas) y las representa en el siguiente mapa del bitexto:



Se concatenan los dos textos y se cruzan. Las áreas oscuras se corresponden con los cruces de los textos en el mismo idioma. La diagonal mayor son las correspondencias entre una palabra y ella misma. Las diagonales menores son más difusas y son las que realmente interesan.

Una búsqueda heurística desvela el camino sobre las diagonales difusas y establece un alineamiento por desplazamiento.

Alineamiento sin usar información léxica

Alineamiento basado en *distancias*: Fung-McKeon

Vector de distancias: Distancias entre las apariciones de una misma palabra.

Por ejemplo si la palabra calidad aparece en las posiciones 10,58,113,181 y 214 del texto, su vector de distancias será:

(48,55,68,33)

Se comparan los vectores de los dos textos, si dos vectores son similares se considera que las palabras pueden coincidir.

Tanto este método como el anterior son apropiados para textos en los que los límites de las frases no están bien identificados.

Índice

- Introducción
- Alineamiento sin usar información léxica
- **Alineamiento usando información léxica**
- Traducción automática

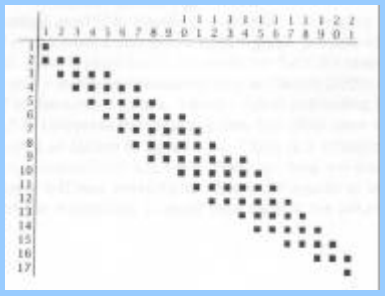
Alineamiento usando información léxica

Método de Kay-Röscheisen

• Asumir que las primeras y últimas frases del bitexto están alineadas (serán las primeras anclas).

• Repetir el siguiente proceso hasta que están alineadas la mayor parte de las frases:

– Construir un conjunto de posibles alineaciones según un criterio de distancia:



– En base a correspondencias léxicas (coocurrencias de palabras) detectar cuáles de las posibles alineaciones son más probables y fijarlas como nuevas anclas.

Alineamiento usando información léxica

Método de Chen

Maximiza la probabilidad del alineamiento, que es considerado una secuencia de *beads*:

$$\arg_A \max \prod_{k=1,\ldots,mA} P(B_k)$$

Es similar al método de los costes de Gale-Church solo que los costes se calculan a través de un modelo de traducción basado en relaciones palabra-palabra.

Alineamiento usando información léxica

Método de Haruno-Yamazaki

Es una variante del método de Kay-Röcheisen pensada para lenguajes muy dispares (como Inglés y Japonés).

- Considera que en lenguajes dispares resulta muy difícil alinear las palabras funcionales (preposiciones, artículos).
- Por tanto, sólo intenta alinear palabras con contenido (verbos, nombres, adjetivos).
- Si los textos son pequeños, no hay contexto suficiente para aplicar las co-ocurrencias del método de Kay-Röcheisen.
- En ese caso propone la utilización de un diccionario para mejorar la identificación de pares de palabras.

Alineamiento usando información léxica

Alineamiento de palabras y expresiones

Un posible resultado adicional de estos métodos es la generación automática de diccionarios bilingües:

• A partir del alineamiento de textos se obtiene un alinemiento de palabras (si es que el método no lo hace al revés).

• Con algún criterio (p.e. frecuencia) se seleccionan las parejas de palabras a incluir en el diccionario.

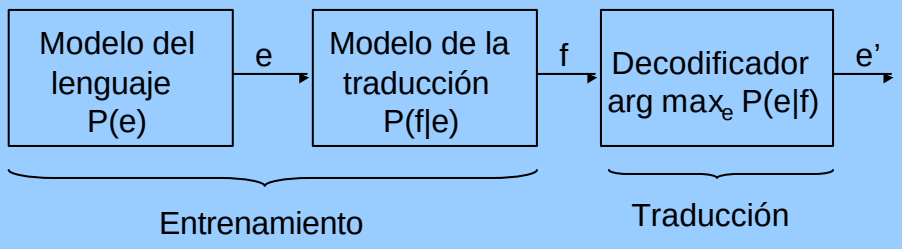
Una versión más general de este problema es la identificación de expresiones y sus correspondientes traducciones.

Índice

Traducción automática

El modelo del canal ruidoso

Para traducir del francés (f) al inglés (e) construimos el siguiente modelo:



$$P(e|f) = \frac{P(f|e)\,P(e)}{P(f)}$$

Traducción automática

El modelo de lenguaje

Nos da la probabilidad de una frase en inglés P(e).
Podemos construir modelos del lenguaje con:

- n-gramas
- gramáticas probabilísticas

Traducción automática

El modelo de traducción

Nos da la probabilidad de una frase en francés f dada una frase en inglés e:

$$P(f|e) = 1/Z \sum_{a1=0..l} \ldots \sum_{am=0..l} \prod_{j=1..m} P(f_j|e_{aj})$$

Suma de todas las posibles alineciones de palabras francesas con palabras inglesas.

l es la longitud de la frase en inglés e

m es la longitud de la frase en francés f

fj es la palabra j-ésima de la frase f

e_{aj} es la palabra de e alineada con f_j

$P(w_f|w_e)$ son las probabilidades de traducción de palabras (estimadas a través de un corpus)

Traducción automática

Decodificador

Aplicando el teorema de Bayes podemos aprovechar las probabilidades calculadas en el entrenamiento para realizar la traducción:

$$e' = \arg_e \max P(e|f) = \arg_e \max (P(e)P(f|e))/P(f) =$$

$$\arg_e \max P(e)P(f|e)$$

El problema es el espacio de búsqueda (generación de frases en inglés e para encontrar la que maximiza la probabilidad).

Podemos utilizar una búsqueda basada en pila para construir estas frases poco a poco.