Instituto Politécnico Nacional

Centro de Investigación en Computación

Laboratorio de Procesamiento
de Lenguaje Natural

Compilación de un corpus paralelo español-inglés alineado a nivel de oraciones

T E S I S

QUE PARA OBTENER EL GRADO DE
MAESTRO EN CIENCIAS DE LA COMPUTACIÓN

P R E S E N T A

Juan Pablo Francisco Posadas Durán

DIRECTORES DE TESIS:
Dr. Grigori Sidorov
Dr. Héctor Jiménez Salazar

MÉXICO, D.F., 2011

Índice general

Índice de figuras

Índice de tablas

Agradecimientos

Agradezco a mis directores de tesis, el Dr. Grigori Sidorov y el Dr. Héctor Jiménez
Salazar por su orientación, apoyo, disposición y atenciones hacia mi persona.

Agradezco a mis sinodales: Dr. Alexander Gelbukh, Dr. Sergio Suárez Guerra, Dr. René García Hernández y Dr. José Luis Oropeza Rodríguez por sus comentarios, sugerencias y observaciones.

Agradezco a mi compañera de toda la vida Andrea Herrera de la Cruz por el apoyo incondicional, palabras de aliento, consejos y paciencia que tuvo conmigo.

Agradezco a mi padre Salomón Posadas Calderón y a mi madre Irene Durán Orozco por el apoyo y la comprensión a lo largo de este trabajo.

Agradezco a mis amigos por su apoyo.

Agradezco al Centro de Investigación en Computación, al Instituto Politécnico Nacional y al CONACyT por las facilidades y recursos otorgados durante mi estancia en el programa de MCC del CIC.

Resumen

Una de las líneas de investigación del Procesamiento de Lenguaje Natural se enfoca en automatizar la alineación de textos paralelos. La utilidad que presenta los textos paralelos alineados es que muestran de manera explícita la relación que existe entre los elementos de un texto en un idioma y los elementos del mismo texto traducido en otro idioma.

En este trabajo de tesis, se plantea un método para la alineación de textos paralelos a nivel de oraciones escritos en los idiomas español e inglés, el cuál utiliza información léxica y estadística bajo un enfoque de programación dinámica. El método utiliza la información léxica contenida en un diccionario bilingüe español–inglés de propósito general restringido (incompleto), así como, el número de elementos significativos y la longitud de la oración medida en términos de caracteres.

El método propuesto se probó en un corpus de textos literarios no balanceados (textos en los que la frecuencia de aparición de alineaciones múltiples, omisiones e inserciones es mayor), en el que reportó una efectividad superior al 90%. Se compararon los resultados obtenidos por el método propuesto contra los obtenidos por el sistema Vanilla aligner (utiliza un enfoque estadístico) utilizando el mismo corpus y se encontró que el método desarrollado fue superior, mostrando un buen desempeño en casos de alineaciones múltiples, omisiones e inserciones.

Por los resultados obtenidos se observa que el uso de la información léxica contenida en un diccionario bilingüe de uso general e información estadística en el método propuesto, hacen de éste un método robusto para realizar la alineación a nivel de oraciones en textos que no presentan una traducción técnica con respecto a métodos exclusivamente estadísticos.

Abstract

Parallel texts alignment is one line of research in Natural Language Processing. The utility of aligned parallel texts is that it shows explicitly the relationship between the elements in a text in one language and elements of the same text translated into another language.

In this thesis, we propose a method for sentence alignment in parallel texts written in Spanish and English, it uses lexical and statistical information in a dynamic programming framework. The lexical information used is the one contained in a bilingual Spanish-English dictionary limited (incomplete) and for general purpose, as well as the sentence length measured in terms of words and in terms of characters.

The proposed method was tested on a corpus of unbalanced literary texts (texts in which the frequency of multiple alignments, omissions and insertions is greater), where we reach a precision aobove the 90%. We compared our results obtained by the proposed method against those obtained by the Vanilla aligner system (which uses a statistical approach)with the same corpus and found that the developed method is superior, particularly in cases of multiple alignments, omissions and insertions.

The results we obtained show that the use of lexical information contained in a bilingual dictionary of general use and statistical information, make this a robust method for sentence alignment in texts that don´t have a technical translation with respect to statistical methods alone.

Capítulo 1
Introducción

En años recientes, cada vez más usuarios a nivel individual o a nivel organización utilizan sistemas de información para almacenar, compartir y administrar su información. Esto se debe a que los sistemas de cómputo han manifestado un incremento significativo en sus capacidades (almacenamiento, procesamiento, comunicación, entre otras), además del desarrollo de programas y servicios orientados al almacenamiento, administración y distribución de la información, que son en gran medida accesibles y amigables.

A la par del aumento en la cantidad de usuarios dentro de los sistemas de información, la cantidad de información disponible en dichos sistemas ha aumentado, esta última con la característica de que la mayor parte de ésta se encuentra expresada en lenguaje natural y en distintos idiomas.

Ante este panorama, el área de Procesamiento de Lenguaje Natural ha propuesto métodos que permiten automatizar el procesamiento de la información expresada en lenguaje natural, buscando reducir el tiempo y esfuerzo humano en el procesamiento y manejo eficiente de dicha información (buscar información relevante sobre un tema en particular y hacerlo lo mas rápido posible). Motivada por la internacionalización de la información (expresada en diferentes idiomas) surge dentro del Procesamiento de Lenguaje Natural una de línea de investigación enfocada a automatizar la alineación de corpus paralelo.

Un corpus paralelo se refiere al conjunto de textos que son traducciones mutuas expresadas en distintos idiomas de un mismo texto. En la Web se puede encontrar varios ejemplos de corpus paralelo: varias empresas transnacionales ponen a disposición de sus usuarios información en varios idiomas sobre sus servicios, políticas, productos, manuales técnicos, entre otros; instituciones de algunos países publican información relevante sobre sus actividades como los parlamentos, secretarías, escuelas, etc.

La alineación de un corpus paralelo consiste en especificar la correspondencia de traducción para cada unidad (párrafo, oración o palabra) en que se dividen los textos que conforman al corpus paralelo. La utilidad que presenta la alineación de textos paralelos es que muestra de manera explícita la relación que existe entre los elementos (párrafo, oración o palabra) de un texto y los elementos del mismo texto traducido en otro idioma.

Un ejemplo de su uso se presenta en sistemas que realizan traducción automática [Brown et al. 1990, Callison-Burch et al. 2006]. En dichos sistemas se requiere de una fase de entrenamiento, en la cuál se obtiene información estadística para realizar la traducción, en estas fases de entrenamiento es donde se utilizan textos

paralelos bilingües alineados ya que en ellos se muestra de manera explicita la relación entre los elementos de los textos en ambos idiomas.

Otras áreas de aplicación para textos paralelos alineados son por ejemplo en la desambiguación de sentidos [Brown et al. 1991a, Chan et al. 2003, Bolshakov et al. 2003], lexicografía bilingüe [Klavans & Tzoukermann 1990,Warwick & Rusell 1990], recuperación de información [Sato 1992], evaluación asistida de traducciones [Pierre 1991] y sistemas para la enseñanza de idiomas [Nerbonne 2000].

1.1. Planteamiento del problema

A pesar de que la capacidad para utilizar el lenguaje natural es algo que todo ser humano posee, éstos utilizan diversos idiomas para comunicarse. Cada idioma tiene una estructura propia que lo hace diferente del resto de tal forma que no existe una traducción directa entre un idioma y otro, por lo que, para realizar una traducción se requiere de cierto conocimiento a priori de los idiomas entre los que se va a traducir.

El aspecto más importante que se toma en cuenta cuando se realiza una traducción es la conservación del contenido; dada la complejidad de los idiomas, la tarea de traducción no es solamente la codificación de la información palabra a palabra, sino a un procesamiento más complejo que toma en cuenta varios elementos (características de los idiomas, contenido del texto original, naturaleza de la información, habilidad del traductor, entre otras) y garantiza la conservación del contenido. Por ejemplo, en ciertas ocasiones suceden adecuaciones por parte del traductor con el fin de que el texto traducido sea entendible, ésto es más frecuentemente en textos literarios donde el traductor puede modificar o eliminar partes de la versión original e inclusive agregar nuevas partes que no se encontraban en la versión original (traducción literaria).

En la alineación de corpus paralelo se trabaja con textos cuya característica principal es que todos son traducciones en diferentes idiomas de un mismo texto en un idioma original y por lo tanto, el tipo de problemas que se deben enfrentar en la alineación de un corpus paralelo incluyen los relacionados con las características estructurales de los idiomas (alfabeto, gramática y léxico) y los relacionados a las adecuaciones realizadas por el traductor (en el caso especifico alineación a nivel de oraciones tenemos las alineaciones múltiples entre oraciones y los casos de inserción u omisión de oraciones).

Para realizar la alineación de corpus paralelo a nivel de oraciones existen tres enfoques: estadístico, estadístico que utiliza información léxica y léxico.
La mayoría de los métodos desarrollados para realizar la alineación de corpus paralelos a nivel de oraciones se basan en el enfoque estadístico debido a su bajo costo computacional y a su capacidad de ser aplicables a varias tuplas de idiomas, sin embargo, presentan la desventaja de que la tasa de error depende de la

similitud entre el corpus de prueba y el corpus a alinear, además, de presentar una baja efectividad en textos con traducción literaria.

Por otro lado, los métodos desarrollados bajo el enfoque estadístico que utiliza información léxica y bajo el enfoque léxico presentan la desventaja de que la mayoría de ellos se han desarrollado para de idiomas que presentan una gran diferencia estructural (por ejemplo, inglés–japonés) y además presentan una baja efectividad en los casos de omisión e inserción de oraciones.

1.2. Justificación

En este trabajo de tesis de se busca desarrollar un método que implemente programación dinámica para encontrar la mejor alineación de un corpus paralelo español–inglés a nivel de oraciones basado en un enfoque estadístico que utilice la información léxica contenida en un diccionario bilingüe español–inglés.

El desarrollo de esta método obedece a la necesidad de contar con un método que disminuya la limitante que implica el tipo de texto en la tarea de alineación evitando así un decremento significativo en la efectividad obtenida al realizar la alineación en textos que presentan una traducción literaria para los idiomas español e inglés. Principalmente lo que se busca con método propuesto es obtener un buen desempeño al realizar la alineación de un corpus paralelo en el que se presenten casos de inserción, omisión y correspondencia múltiple entre oraciones.

El método propuesto se centra en el par de idiomas español e inglés por dos razones: la primera razón es la falta de herramientas para la alineación de corpus paralelo español–inglés a nivel de oraciones que sea robusto en textos que presenten una traducción literaria y la segunda razón es que el inglés es uno de los idiomas más utilizados a nivel mundial por lo que es fácil encontrar corpus paralelo en español e inglés.

Se eligió evaluar el método propuesto con un corpus de textos literarios debido a que éstos presentan una traducción literaria y por lo tanto los casos de inserción, omisión y alineación múltiple tienen una frecuencia de ocurrencia mayor en este tipo de textos a diferencia de algunos otros como: manuales técnicos, artículos, entre otros. Se plantea como trabajo futuro el realizar pruebas con distintos tipos de textos.

En el futuro, el desarrollo del método propuesto proveerá una gama de opciones en el desarrollo de nuevas herramientas de procesamiento de información expresada en lenguaje natural y en los idiomas español–inglés, que requieran de una alineación a nivel de oraciones en alguna de sus etapas sin que el tipo de corpus a ser alineado sea un factor.

1.3. Hipótesis de trabajo

La hipótesis de trabajo que se plantea en esta tesis es la siguiente: mediante el uso de un diccionario bilingüe español–inglés es posible agregar cierto nivel de certidumbre que ayude en la tarea de alinear las oraciones de un corpus paralelo bilingüe español–inglés tal que la efectividad en la alineación no disminuya considerablemente al probarlo en un corpus paralelo en el que la traducción no sea técnica (textos literarios). Se asume que el uso de un diccionario bilingüe provee una fuente de información en la que se establece la correspondencia entre palabras de ambos idiomas.

1.4. Objetivos

A continuación se describen el objetivo general y los objetivos particulares de este trabajo.

1.4.1. Objetivo general

Desarrollar un método para la alineación de un corpus bilingüe español–inglés a nivel de oraciones utilizando un diccionario bilingüe inglés–español e implementarlo para compilar un corpus bilingüe español–inglés de textos literarios (los cuales presentan los casos más difíciles de alineación) alineado a nivel de oraciones.

1.4.2. Objetivos particulares

- Diseñar un método para la alineación del corpus a nivel de oraciones.
- Implementar el método de alineación propuesto.
- Aplicar el método de alineación en el corpus de prueba.
- Registrar el nivel de efectividad logrado por el método.
- Comparar la efectividad obtenida por el método propuesto y la obtenida por un método estadístico (particularmente se utilizó el sistema Vanilla Aligner).
- Compilar un corpus paralelo español–inglés alineado a nivel de oraciones.

1.5. Alcances

Los beneficios derivados de esta propuesta de tesis van dirigidos principalmente a investigaciones del área de Procesamiento de Lenguaje Natural y áreas afines, que se interesen en el desarrollo de herramientas enfocadas en los idiomas español e inglés cuya unidad de trabajo sea o dependa de la oración.

Con el desarrollo e implementación del método propuesto se espera como beneficios:

- Desarrollar una herramienta de alineación de corpus paralelo español–inglés a nivel de oraciones.
- Compilar un corpus paralelo español–inglés alineado a nivel de oraciones.
- Utilizar el método propuesto como una opción para comparar contra métodos similares.
- Utilizar el método propuesto en el desarrollo de nuevas herramientas que requieran de una alineación de corpus paralelo español–inglés a nivel de oraciones en alguna de las siguientes áreas:

- Alineación de corpus paralelo a nivel de palabras o un nivel de de detalle mayor:
    - Alineación de corpus paralelo para más de dos idiomas.
    - Traducción automática.
    - Desambiguación de sentidos.
    - Sistemas de enseñanza de idiomas.
    - Lexicografía bilingüe.
    - Sistemas de recuperación de información.

Capítulo 2
Marco teórico

2.1. Alineación de textos paralelos

2.1.1. Introducción

Existen diversos idiomas los cuales se asocian generalmente a la nacionalidad de sus usuarios. El origen y evolución de los idiomas, así como su presente, son temas de investigación dentro de la lingüística.
La definición de un idioma no es más sencilla que la definición de lenguaje, aunque podemos decir que corresponde a una caracterización particular de éste utilizada por un determinado número de individuos. Al estar relacionados con los seres humanos, los idiomas son también dinámicos en el tiempo y se encuentran determinados por factores sociales, económicos, políticos, culturales, biológicos, históricos, etc.

Un fenómeno fácilmente comprobable, que se presenta en los idiomas es que dos idiomas no son equivalentes entre si al mismo tiempo, es decir, el hecho de que una persona domine determinado idioma no implica en automático que domine algún otro idioma. Algunas de las teorías sobre el origen de los idiomas concuerdan con la existencia de un origen común, que por diversos factores, se fue transformando y diversificando gradualmente a medida que los seres humanos comenzaron a extenderse por el planeta [Oliphant 1996, Kirby & Hurford 1997], aunque esta teoría explica que algunos idiomas tengan similitudes estructurales, en la actualidad dos idiomas son lo suficientemente diferentes estructuralmente como para que no se dominen ambos con solo conocer uno de los dos.

Este fenómeno que presentan los idiomas hace necesario recurrir a la traducción entre los idiomas. Se define a la acción de traducir como[1]:

“Expresar en una lengua lo que esta escrito o se ha expresado antes en otra”.

Típicamente la tarea de traducción requiere de conocer ambos idiomas (gramática, léxico y semántica de cada idioma), tener conocimiento del área a la que pertenecen los textos (tecnicismos y acepciones de las palabras) e inclusive conocer algo de la cultura de los países hablantes a los que pertenecen esos idiomas, lo que la convierte en una tarea no trivial.

[1] Véase Diccionario de la Real Académia de la Lengua Española, Junio 2011, http://buscon.rae.es/draeI/SrvltConsulta?TIPO_BUS=3&LEMA=traducir.

La tarea de traducción de textos, de manera general, se puede realizar de dos formas dependiendo del tipo de texto a traducir: traducción literaria y traducción técnica.
La traducción literaria es la traducción a algún tipo de trabajo literario, mientras que la traducción técnica es la traducción de textos especializados en alguna área de interés [Trujillo 1999].

Una diferencia importante entre ambos tipos de traducción radica en el hecho de que en la traducción técnica la conservación del contenido es generalmente el criterio de traducción más importante, mientras que en la traducción literaria no necesariamente se conserva el contenido, es decir, se pueden agregar partes nuevas que no existan en el texto original o se pueden eliminar partes que existan en el texto original según criterio del traductor o del autor [Trujillo 1999].

Se puede plantear la tarea de alineación de textos paralelos como una necesidad para desarrollar la tarea de traducción de un texto de un idioma a otro; de acuerdo con este enfoque, la necesidad de alineación de textos paralelos surge con el planteamiento de la siguiente pregunta: ¿qué información se requiere saber para automatizar la tarea de traducción entre idiomas?. La tarea de alineación de textos paralelos resuelve esta pregunta al exponer la relación entre los elementos de un texto en un idioma y los elementos de un texto en otro idioma [Véronis 2000].

2.1.2. Características de la alineación de textos paralelos

La tarea de alineación de textos paralelos la definimos como sigue:

Sea A un texto en un idioma Z (idioma origen) dividido en n unidades definidas por el usuario y sea B el mismo texto traducido en un idioma X (idioma destino) dividido en m unidades, con m y n no necesariamente iguales, entonces, la acción de alinear los textos consiste en determinar la correspondencia entre las unidades de un texto A y las unidades del texto B.

A la unidad que se utilice para dividir a los textos paralelos se le llama nivel de detalle o granularidad, estas unidades van desde las más generales hasta las mas particulares como son: capítulo, sección, oración o palabra respectivamente.

Típicamente se trabaja en alguno de los siguientes tres niveles de detalle: alineación a nivel de párrafos, alineación a nivel de oraciones y alineación a nivel de palabras [Brown et al. 1991b].

Los niveles de detalle más particulares se relacionan implícitamente con los niveles de detalle mas generales, por ejemplo, el nivel de detalle de palabras implica el conocimiento de las oraciones en las que se encuentran dichas palabras, así como el párrafo en el que se encuentra dicha oración, en otras palabras, en cualquiera de las presentaciones en que se maneje el texto, se observa que la alineación de textos paralelos no solo depende de la similitud entre

unidades, sino también influye la posición relativa en el texto de cada unidad [Véronis 2000].

En el caso particular de la alineación a nivel de oraciones, se presentan ocho casos de correspondencia entre las oraciones de un corpus paralelo (al texto en el idioma original se le llama texto en el idioma origen y a su traducción de le llama texto en el idioma destino), los cuales se mencionan a continuación [Simard et al. 1992]:

Una oración en el idioma origen corresponde solamente a una oración en el idioma destino (1:1).
Una oración en el idioma origen corresponde a dos oraciones en el idioma destino (1:2).
Dos oraciones en el idioma origen corresponden a una oración en el idioma destino (2:1).
Una oración en el idioma origen no tienen correspondencia en el idioma destino (1:0) (caso de omisión).
Una oración en el idioma destino no tienen correspondencia en el idioma origen (0:1) (caso de inserción).
Dos oraciones en el idioma origen corresponden a dos oraciones en el idioma destino (2:2).
Tres oraciones en el idioma origen corresponden a una oración en el idioma destino (3:1).
Una oración en el idioma origen corresponde a tres oraciones en el idioma destino (1:3).

En la tabla 2.1 se muestran algunos ejemplos de los tipos de correspondencia entre oraciones [2].

Tabla 2.1: Ejemplos de tipos de alineaciones

Alineación	Texto en inglés	Texto en español
1:1	To Sherlock Holmes she is always the woman	Para Sherlock Holmes, ella es siempre la mujer
1:2	He was a remarkably handsome man, dark, aquiline and moustached, evidently the man of whom I had heard	Se trataba de un hombre bien parecido, moreno, de nariz aguileña y con bigote. Evidentemente, el mismo hombre del que había oído hablar.
2:1	I told you that I would call. He looked from one to the other of us, as if uncertain which to address.	Le dije que vendría a verle –nos miraba a uno y a otro, como si no estuviera segura de a quién dirigirse.

Tabla 2.1: Ejemplos de tipos de alineaciones (continuación)

Alineación	Texto en inglés	Texto en español
3:1	It would be injustice to hesitate, said he. You will, however, I am sure, excuse me for taking an obvious precaution. With that he seized my hair in both his hands, and tugged until I yelled with the pain.	Sería una injusticia dudar de usted –dijo–, pero estoy seguro de que me perdonará usted por tomar una precaución obvia –y diciendo esto, me agarro del pelo con las dos manos y tiró hasta hacerme chillar de dolor–.
1:3	Why, dear me, it sounds quite hollow!, he remarked, looking up in surprise.	Pero... ¡Valgame Dios! ¡Esto suena hueco! –exclamó, alzando sorprendido la mirada.
2:2	I leave a photograph which he might care to possess and I remain dear Mr. Sherlock Holmes. Very truly yours	Dejo una fotografía que tal vez le interese poseer. Y quedo, querido señor Sherlock Holmes, suya afectísima.

Los últimos cinco casos (1:0, 0:1, 2:2, 3:1, 1:3) presentan una frecuencia de aparición muy pequeña, por lo que generalmente se trabaja con los tres primeros casos (1:1, 1:2, 2:1). De éstos, el caso más simple corresponde a aquel en el que cada oración en el idioma origen corresponde solamente a una oración en el idioma destino de manera secuencial dentro de un párrafo y cada una de las palabras que componen a la oración en el idioma origen corresponden a una palabra en la oración en el idioma destino, a este caso particular se le llama traducción literal.

Por otro lado, el caso más difícil se presenta cuando la correspondencia entre las oraciones en el idioma origen y las oraciones en el idioma destino es múltiple o no existe correspondencia (casos de omisión e inserción).

En una alineación de corpus paralelo se asume que el corpus a alinear no presenta casos de referencias cruzadas, es decir, se asume que se crea el texto traducido siguiendo el orden que presentan las oraciones en el texto original. Sin embargo, existen casos en los que no necesariamente la versión original y la versión traducida tienen el mismo orden ya que es posible la adición u omisión de oraciones en la versión traducida, este fenómeno se atribuye directamente a la persona que realiza la traducción [Gelbukh & Sidorov 2006a, Véronis 2000].

En el desarrollo de métodos para alineación se requiere contar con un corpus de prueba, cualquier par de textos que sean traducciones mutuas alineados, sin importar su tipo o contexto, pueden ser utilizados; sin embargo, el principal inconveniente que se tiene es la disponibilidad de los mismos: se requiere de textos diversos tipos (técnicos, literarios, institucionales, científicos, etc.) y

traducidos en ciertos idiomas de interés, cada par de textos (idioma origen y destino) debe de incluir su alineación (codificada de alguna manera), realizada por un conjunto de personas especializadas en traducción que indiquen los criterios utilizados para la alineación de los textos y que además se encuentren disponibles en formato electrónico [Véronis & Langlais 2000].

A pesar de la escasa disponibilidad de textos alineados, es posible contar con conjuntos de textos paralelos a través de la Web, que pueden ser alineados manualmente para la conformación de un corpus de prueba, por otro lado, existen distintos grupos de investigación que han desarrollado proyectos enfocados a la compilación de corpus alineados para diferentes idiomas [Véronis & Langlais 2000, Singh et al. 2000, Isahara & Haruno 2000].

Los textos que usualmente han sido utilizados en sistemas dedicados a la alineación de textos se pueden clasificar de la siguiente manera [Véronis & Langlais 2000]:

- Textos institucionales: comprende debates parlamentarios, reportes oficiales, documentos legales, etc. Son el tipo de texto más utilizado por dos razones: la primera es que su traducción es técnica, es decir, su alineación se compone principalmente de casos del tipo 1:1 (debido a que los traductores no toman riesgos en la traducción) y la segunda razón es que se encuentra en grandes cantidades. Ejemplos de este tipo de textos son los procesos judiciales del Parlamento de Canadá y reportes bancarios de Suiza.
- Manuales técnicos: este tipo de textos contienen una mayor concentración de términos técnicos, una estructura compleja (tablas, figuras, glosario, índices, etc.) y se espera que la traducción sea técnica.
- Textos científicos: este tipo de textos contienen términos técnicos y una estructura compleja, sin embargo, a diferencia de los manuales técnicos el traductor tiene mayor libertad en su traducción, especialmente donde se necesite una adaptación que ayude a la argumentación del autor.
- Literatura: en este tipo de textos la traducción rara vez se realiza de manera técnica y son más frecuentes la adaptaciones a los elementos del texto por parte del traductor (por ejemplo omisión e inserción).

2.2. Técnicas de optimización

2.2.1. Introducción

Cuando resolvemos algún problema generalmente lo que se hace es buscar una solución que satisfaga las condiciones propias del problema a resolver, para esto se construye un modelo que nos permita buscar la o las soluciones de un conjunto de posibles soluciones (todas aquellas que satisfacen parcial o completamente las

restricciones del problema). Al conjunto de posibles soluciones del problema se le llama espacio de soluciones, espacio de búsqueda o espacio del problema y a cada uno de los elementos del conjunto se le llama solución candidata [Jones 2008].

Generalmente se suele pensar en el espacio de soluciones en términos del número de acciones que se pueden realizar para ir desde un punto inicial hasta un punto final y de los puntos intermedios que se generaron al efectuar dichas acciones. A estos puntos intermedios se les llama estados [Michalewicz & Fogel 1998].

Existen tres elementos que deben de ser claros al momento de resolver cualquier problema [Michalewicz & Fogel 1998]:

- La representación: se refiere a la codificación aplicada a las soluciones candidatas para que estas puedan ser manipuladas. La representación utilizada para resolver un problema en particular y su interpretación, determinan el espacio de búsqueda y su tamaño.
- El objetivo: describe el propósito que se esta buscando, generalmente se utiliza una expresión matemática que defina la tarea que se debe de cumplir.
- La función de evaluación: es un mapeo del espacio que contiene a las posibles soluciones hacia un conjunto de números, donde a cada posible solución se le asigna un valor numérico que indique la calidad de la solución; estas pueden ser funciones de evaluación ordinarias(aquellas funciones de evaluación permiten distinguir solamente una graduación de todas las posibles soluciones) y funciones de evaluación numérica (no solamente asignan una graduación sino también indican el nivel de calidad de la solución candidata). Para elegir o diseñar una función de evaluación es recomendable tomar en cuenta los siguientes aspectos:

    1. Cuando una solución satisface completamente el objetivo, ésta debe ser evaluada con la mejor calificación.

    2. Algunas veces las soluciones de interés se encuentran en un subconjunto del espacio de búsqueda, la función debe de tomar en cuenta las restricciones específicas del problema.

2.2.2. Problemas de optimización

Para resolver un problema de otpimización se utiliza un modelo matemático, en el cual se representen la forma en que los elementos del problema se relacionan entre sí.

Los componentes básicos de este modelo son [Nemhauser 1966]:
1. Las variables: estos elementos pueden ser manipulados para alcanzar el objetivo.

2. Los parámetros: estos elementos se encuentran presentes en el problema pero no pueden ser controlados.
3. La función de evaluación : es el valor que se asocia a valores particulares de de las variables y parámetros.
4. La región de factibilidad: esta región se encuentra especificada por límites o restricciones en los valores que las variables pueden tomar, generalmente es posible representar parte de los límites o restricciones utilizando ecuaciones o desigualdades.

Una solución cuyas variables se encuentren en la región factible se le llama solución factible y al conjunto de soluciones factibles se le llama parte factible o espacio factible.

Un problema de optimización o problema de búsqueda es aquel cuyo objetivo es encontrar la mejor solución factible en un espacio de búsqueda dado. A continuación se da una definición formal de un problema de optimización [Michalewicz & Fogel 1998]:

Definición 2.1 *Dado un espacio de búsqueda S junto con su parte factible $F \subseteq S$, encontrar $x \in F$ tal que*

$$eval(x) \leq eval(y)$$

para toda $y \in F$.

La solución x que satisface la condición mencionada en la definición 2.1 se le llama solución global u óptimo.

2.2.3. Algoritmos de optimización

Debido a que las restricciones de un problema cambian de problema en problema, no existe un algoritmo general de optimización, en cambio se han desarrollado diversos algoritmos de optimización que son aplicables a un conjunto amplio de problemas en específico.

Existen varios enfoques para clasificar a los algoritmos de optimización tomando en cuenta la forma en que opera cada uno de ellos. Una forma de clasificarlos es de acuerdo a como evalúan las soluciones candidatas, teniendo así dos clases [Michalewicz & Fogel 1998]:

- Algoritmos que solo evalúan soluciones completas.
- Algoritmos que requieren de la evaluación de soluciones parcialmente construidas. Generalmente estas se presentan de dos formas:
    - Como una solución incompleta del problema original.

- Como una solución completa de una versión simplificada del problema.

Ejemplos de algoritmos de cada una de las clases se presentan en la tabla 2.2 y en la tabla 2.3 se mencionan algunas de las ventajas y desventajas que presentan cada una de las clases.

Tabla 2.2: Ejemplos de algoritmos de optimización

Algoritmos que trabajan con soluciones completas	**Algoritmos que trabajan con soluciones parciales**
Búsqueda exhaustiva	Programación dinámica
Búsqueda tabú	Dividir y conquistar
Hill Climbing	Algoritrmos codiciosos (greedy)
Algoritmos evolutivos	Búsqueda en profundidad
Recocido simulado	Algoritmo A*

Tabla 2.3: Ventajas y desventajas de los tipos de algoritmos

	Algoritmos que trabajan con soluciones completas	**Algoritmos que trabajan con soluciones parciales**
Ventajas	En cualquier momento se tiene una solución potencial	Se puede descomponer el problema original en subproblemas más fáciles de resolver
Desventajas	Se requiere conocer todas las variable de decisión	Idear una forma de organizar los subespacios de los subproblemas. Se requiere por lo menos dos funciones de valuación: una para evaluar las solución completa y otra para evaluar la calidad de las soluciones parciales.

2.2.4. Programación dinámica

La programación dinámica es un algoritmo de optimización que utiliza un enfoque de descomposición del problema en una serie de subproblemas más pequeños, lo cuales al ser resueltos pueden combinar cada una de las soluciones para obtener la solución al problema original.

La programación dinámica es un procedimiento recursivo, en el que cada subproblema o estado intermedio es una función de los subproblemas o estados ya resueltos. Los pasos del algoritmo son [Nemhauser 1966]:

1. Descomponer el problema original en una serie de subproblemas $Q_1, \ldots, Q_N$ de los que se encontrará el valor óptimo $f_N(X_N)$, $f_{N-1}(X_{N-1})$, . . . , $f_1(X_1)$ de la siguiente manera:

$f_1(X_1) = max\ Q_1(X_1, D_1) = max\ r_1(X_1, D_1)$

...

$f_N = max\ Q_N(X_N, D_N) = max\ [r_N(X_N, D_N) + f_{N-1}(t_N(X_N, D_N))]$

donde:

X es el conjunto de parámetro para un subproblema en especifico.

D es el conjunto de variables que se modifican para tomar una decisión.

$r()$ es una función escalar que asigna una evaluación a un conjunto particular de variables y parámetros.

$t()$ es una transformación que regresa un solo valor, la cual expresa el resultado de cada estado en función de los parámetros y variables

2. Encontrar la mejor solución de cada subproblema hasta llegar al subproblema final (N), el cual contiene la respuesta al problema original.

Los problemas prototipo que se pueden resolver utilizando este método tienen las siguientes propiedades [Michalewicz & Fogel 1998]:

- Debe ser posible descomponer el problema en una secuencia de decisiones hechas en varias etapas.
- Cada etapa debe tener un número determinado de posibles estados.
- Una decisión lleva de un estado en una etapa hacia algún estado en las siguiente etapa.
- Existe un costo bien definido al viajar de un estado a otro a través de las etapas.

Capítulo 3
Estado del arte

3.1. Introducción

No se sabe la fecha precisa en que se comenzó a trabajar por primera vez en la alineación de textos multilingües paralelos. Es claro que no pudo comenzar hasta el momento en que existieron por lo menos dos idiomas, momento que también es difícil de determinar en el tiempo.
A medida que los seres humanos comenzaron a poblar el planeta y que los idiomas comenzaron a diversificarse, surge naturalmente la necesidad de realizar traducciones de textos; por ejemplo, podemos situarnos en los tiempos de las primeras grandes civilizaciones como la Romana (753 a.C.- 478 d.C.) la cual extendió sus dominios por gran parte de Europa, Asia y África y la necesidad de comunicarse entre su lengua (latín) y la lengua hablada en cada una de las regiones conquistadas. Con base en dicha necesidad, puede pensarse que los textos paralelos alineados eran utilizados como recursos lingüísticos para la enseñanza de los idiomas, ya que esta técnica permitía generar listados de correspondencia entre los símbolos empleados en los idiomas y utilizarlas como una guía [Véronis 2000].

Uno de los primero casos de estudio de interés para el área de alineación de corpus paralelo, fue aquel relacionado con el desciframiento de la escritura egipcia en 1822. Este hecho sobresale porque se trata del estudio de la lengua antigua de una de las civilizaciones importantes y que nadie podía descifrar hasta entonces.

El elemento que jugó un papel importante en la solución del problema fue la llamada piedra Rosetta (Figura 3.1), encontrada en 1799 por soldados franceses del fuerte Julien en el pueblo de Rosetta a orillas del río Nilo. La característica que posee dicha piedra y que la hace tan importante es que en ella se encuentra inscrito un mismo texto en tres escrituras diferentes: griego,demótico y jeroglíficos. Los jeroglíficos y el demótico son dos escrituras diferentes de la misma lengua egipcia, pero con pequeñas diferencias en los símbolos utilizados, tal que la primera se utilizaba más con fines decorativos y la segunda para escritos [Singh 1999].

El objetivo era descubrir el significado de los símbolos utilizados por los egipcios tanto en la escritura demótico como en los jeroglíficos, a través de la alineación de los textos en egipcio con el texto en griego, que era conocido. El problema no era fácil debido a que la piedra Rosetta no se logró recuperar de manera íntegra.

Fue en 1822 que un lingüista francés llamado Jean Francois Champollion descifró la escritura egipcia, encontrando la correspondencia entre ésta y la escritura griega.

Champollion partió del supuesto de que la escritura egipcia era fonética y comenzó a hacer la correspondencia entre idiomas, en primera instancia de la representación de nombres propios en ambos idiomas buscando en la piedra Rosetta y algunos pergaminos egipcios, para finalmente intentar definir la correspondencia de aquellos símbolos con base en la piedra Rosetta.

Figura 3.1: *La piedra Rosetta, tiene en la parte superior el texto inscrito en jeroglíficos, en la parte media el texto inscrito en demótico y en la parte baja el texto inscrito en griego*

Durante la Segunda Guerra Mundial (1939-1945), se construyeron máquinas que realizaban las tareas de desciframiento y de encriptación de textos de manera semiautomática.

Posteriormente, ya con el desarrollo de la primer computadora y ante el potencial mostrado por las máquinas en las tareas de desciframiento y encriptación, WarrenWeaver escribe un memo en 1955 en el que se especifican las primeras técnicas para implementar la traducción entre distintos idiomas en las computadoras y a partir de entonces comienza la investigación en lo que se conoce como la traducción automática por computadora (Machine Translation MT), una de las primeras áreas de investigación dentro de la Lingüística Computacional y también una en la que más investigación se ha realizado [Weaver 1955].

En la década de los cincuenta y sesenta, investigadores en el área de MT proponen el uso de ciertos recursos de información que ayuden al desarrollo de nuevas metodologías, como por ejemplo, traductores especializados, diccionarios bilingües, listado de frases propias de los idiomas, entre otros.

Existen algunos trabajos [King 1956, Koutsoudas & Humecky 1957, Hays 1963] en los que se propone utilizar textos paralelos multilingües alineados como un recurso de información, ya que de ellos se obtendría información estadística sobre la correspondencia entre palabras de un idioma y sus traducciones en otro idioma, lo que ofrecería algunas pistas para realizar la traducción.
En las secciones subsecuentes se comentará sobre las investigaciones que se han realizado en el tema y especialmente de las metodologías que se han propuesto para automatizar la alineación de corpus paralelo.

3.2. Alineación automática de corpus paralelo

Aunque a finales de la década de los 50 y principios de la década de los 60, ya se había sugerido el uso de corpus paralelo alineado como recurso lingüístico en las metodologías de traducción automática, no se publicaron trabajos que hicieran uso de éstos debido a que no eran muy accesibles: no existía un compendio de corpus previamente alineados y la generación de un corpus alineado requería de especialistas que realizaran la alineación del corpus para un par de idiomas deseado. Se pensó en la idea de automatizar la tarea de alineación de corpus paralelo, sin embargo, especialistas consideraban que la capacidad de cómputo que se tenía era limitada y decidieron posponer las investigaciones al respecto.

Fue en el año de 1991 que se publican las primeras metodologías para automatizar la tarea de alineación de textos paralelos a nivel de oraciones y en los años posteriores se comenzaron a publicar nuevas metodologías que extendían su nivel de detalle a las palabras.

En la actualidad, las metodologías enfocadas a la alineación a nivel de oraciones existentes en la literatura se pueden clasificar en dos tipos: metodologías estadísticas y metodologías estadísticas que usan información léxica. En las siguientes secciones se discuten algunos ejemplos de interés de ambos enfoques y en la tabla 3.1 se muestra un resumen de los resultados obtenidos por dichos ejemplos.

Tabla 3.1: Comparativo de metodologías propuestas

Característica	Gale & Church [1991]	Brown et al. [1991b]	Gautam & Sinha [2007]
Tipo de metodología	Estadística	Estadística	Estadística con información léxica
Procedencia del corpus	U. Bancos Suizos	Parlamento Canadiense	Literario
Idiomas de trabajo	inglés, francés y alemán	inglés y francés	inglés e hindi
Característica base	número de caracteres	número de palabras	número de palabras, caracteres y traducciones mutuas
Tasa de error	2 %	2.3 %	20 %

3.2.1. Metodologías estadísticas

En las metodologías estadísticas se busca encontrar una configuración particular de oraciones alineadas, el cual llamaremos alineación y denotaremos por A, que sea la que tenga mayor probabilidad de ocurrir en los textos paralelos O y T con base en un modelo del lenguaje propuesto:

$$max P\,(A|O,T)\ \ equivale\ a\ max P(A,O,T) \qquad (3.1)$$

Una alineación A se puede descomponer en una secuencia de cuentas $(B_1, \ldots, B_k)$ y suponer que la probabilidad de cada una de las cuentas depende solamente de las oraciones que la conforman y que es independiente de las demás cuentas, así podemos aproximar la probabilidad de la alineación con base en la probabilidad de las cuentas como:

$$P(A,O,T) \approx \prod_{i=1}^{k} P(B_i) \qquad (3.2)$$

Ahora el problema se reduce en encontrar una forma de estimar las probabilidades de cada una de las cuentas, basándose en las oraciones que conforman cada una de las cuentas.

Con este enfoque, la metodologías dividen el problema de alineación de manera general en dos subproblemas:

- El primer subproblema consiste en encontrar una función de evaluación que describa la correspondencia entre las oraciones que conforman al corpus paralelo.
- El segundo subproblema consiste en encontrar la mejor alineación posible entre las oraciones del corpus.

El primer subproblema generalmente requiere de un análisis del corpus a alinear, en el cual se selecciona una o unas características que se pueden cuantificar del corpus que sean significativas para realizar la alineación, como por ejemplo, relación entre el número de palabras que conforman las oraciones en cada uno de los idiomas o la posición relativa que guardan las oraciones que corresponden entre sí.

Una vez definidas la o las características sobre las que se va a trabajar, se procede a realizar una fase de exploración o de entrenamiento, en la cual se obtienen cierta información estadística (información empírica) sobre una porción pequeña del texto paralelo de prueba. Posteriormente se utiliza esta información para asociar el comportamiento de la o las características seleccionadas a una distribución de probabilidad ya conocida (modelo del lenguaje).

Se espera que la función de evaluación propuesta haga perceptible la diferencia entre aquellas parejas que deben ser alineadas y aquellas que no. Se propone un

elemento especial llamado «vacio» para evaluar los casos en los que una oración no tiene correspondencia.

A este tipo de metodologías también se le conoce como metodologías basadas en información interna porque se basan solamente en la información que se puede obtener de los texto y no requieren de ninguna información lingüística sobre los idiomas en cuestión.

El segundo subproblema consiste en encontrar la mejor alineación posible de un corpus paralelo en particular, que se obtiene con base en la alineación de los elementos que conforman el corpus (recordemos que la correspondencia no necesariamente es univoca, sino que se debe permitir correspondencia múltiple e inclusive correspondencia con el elemento especial «vacio»). Típicamente se lleva cabo implementando un algoritmo que verifique todas las posibles alineaciones para cada una de las oraciones en un idioma en particular, utilizando el enfoque de programación dinámica.

En ocasiones estas metodologías se auxilian de elementos clave que sugieran una correspondencia entre un par de oraciones en el corpus llamados anclas (anchor points), estos elementos se utilizan para delimitar la búsqueda de posibles asociaciones para una oración determinada y evitar la propagación de errores; ejemplos de anclas son cognados (cognate), fechas, números, signos de puntuación, límites de un párrafos, títulos, entre otros. La anclas utilizadas dependen del tipo de corpus que es utilizado y su selección generalmente es estimada debido a que no se conoce información detallada sobre la correspondencia entre ambos textos de antemano.

De manera general lo que arrojan como resultados estas metodologías puede ser una relación de las posiciones de las oraciones en un texto que corresponden a las posiciones de las oraciones en la traducción del texto o un arreglo que contenga a las oraciones alineadas o cuentas.

Algunos de los primeros trabajos de este tipo que tratan el problema de la alineación de textos paralelos a nivel de oraciones con un enfoque estadístico fueron realizados por Brown et al. [1991b]; Gale & Church [1991] y Wu [1994]. A continuación describiremos de manera general las características de algunas metodologías propuestas y los resultados obtenidos.

Gale & Church (1991)

En este trabajo se describe un método de alineación basado en un modelo estadístico sobre la longitud de las oraciones medida en caracteres, cuya idea central es que una oración larga (medida en caracteres) en un idioma tiende a ser traducida en una oración larga en el otro idioma, y de forma análoga, una oración corta en un idioma tiende a generar una oración corta en el otro idioma. Se asigna una puntuación a cada posible cuenta con base en la diferencia de longitudes entre las oraciones y a la varianza de esta diferencia.

Se limitan los casos de estudio de las posibles alineaciones entre oraciones a los siguientes seis casos:

- Una oración en el texto O corresponde a una oración en el texto T(relación 1:1).
- Una oración en el texto O corresponde a dos oraciones en el texto T(relación
- 1:2).
- Dos oraciones en el texto O corresponden a una oración en el texto T(relación
- 2:1).
- Dos oraciones en el texto O corresponden a dos oraciones del texto T(relación
- 2:2).
- Una oración en el texto O es eliminada en el texto T(relación 1:0).
- Una oración que no estaba en el texto O es agregada en el texto T(relación 0:1).

La alineación a nivel de oraciones se lleva a cabo en dos etapas:

1. En la primer etapa se lleva a cabo la alineación a nivel de párrafos de los textos O y T. Se define a un párrafo como un bloque de caracteres que tienen un longitud mayor a 100 caracteres y se asocian con base en el orden en que aparecen en los textos respectivamente; los bloques de caracteres que no sean párrafos se ignoran.
2. En la segunda etapa se lleva acabo la alineación de las oraciones en cada uno de los párrafos alineados.

La alineación de las oraciones en cada uno de los párrafos se realiza con ayuda de una implementación basada en programación dinámica, que busca la alineación A conformada por las cuentas que presenten la mayor puntuación asignada con respecto al resto de las posibles cuentas(restringido a los casos de estudio antes mencionados).

La puntuación se asigna de acuerdo con el siguiente modelo del lenguaje:

- Se asigna una puntuación a cada una de las cuentas de acuerdo con la siguiente expresión

$$D(B_1) = -logP(match|\delta) \tag{3.3}$$

donde $match$ es alguno de los seis casos posibles de correspondencia (1:1,1:2,...) y δ es una variable que depende de las longitudes de las oraciones que conforman la cuenta.

- Se definen variables aleatorias que indican el número de caracteres que genera un carácter del texto O en el texto T. Se considera que estas variables son independientes e idénticamente distribuidas, y que son modeladas por una distribución normal con media μ (número de caracteres esperados en T por carácter en O) y varianza s^2(varianza del número de caracteres en T por carácter en O).

- δ compara la diferencia en la suma de las longitudes de las oraciones (l_1, l_2) de la cuenta con la media y la varianza de todo el corpus:

$$\delta = (l_2 - l_1\mu) / \sqrt{l_1 s^2} \tag{3.4}$$

- Utilizando el teorema de Bayes para determinar la probabilidad de la ecuación 3.3, obteniendo:

$$P(match|\delta) = P(match)P(\delta|match) \tag{3.5}$$

- $P(match)$ y $P(\delta|match)$ se obtienen de manera experimental del corpus de prueba.

El corpus utilizado para los experimentos fueron los reportes económicos emitidos por la Unión de Bancos de Suiza(Union Bank of Switzerland, USB), los cuales se encontraban disponibles en tres idiomas: inglés, francés y alemán.
De manera general el algoritmo obtuvo un error del $4{,}2\%$ sobre un total de 1326 alineaciones en los corpus de prueba inglés–francés e inglés–alemán. El algoritmo presenta un buen desempeño para las alineaciones 1:1 en las que presenta una tasa de error del 2%y tiene muchas dificultades para los casos 1:0 y 0:1 en los que presenta una tasa de error del 100%, en la tabla 3.2 se muestran las tasas de error para los seis casos posibles de alineación en el total de alineaciones(ambos corpus de prueba).

Tabla 3.2: Tasa de error por alineación

Tipo de alineación	Numero de alineaciones	Error	Porcentaje
1–0	13	13	100 %
1–1	1167	23	2 %
2–1	117	10	9 %
2–2	15	5	33 %
3–1	2	2	100 %
3–2	1	1	100 %

Brown et al. (1991)

En este trabajo se describe un método de alineación basado en un modelo estadístico sobre la longitud de las oraciones medida en palabras (tokens), cuya idea central es que las longitudes de las oraciones que están relacionadas presentan una alta correlación, es decir, una oración larga (medida en palabras) en un idioma tiende a ser traducida en una oración larga en el otro idioma, y de forma análoga en el caso de una oración corta.

El corpus utilizado en los experimentos fueron los reportes judiciales emitidos por el Parlamento canadiense que son referidos como Hansards, los cuales se encontraban disponibles en dos idiomas: inglés y francés.

A diferencia de Gale & Church [1991] se asume que cada corpus es una secuencia de oraciones que incluyen delimitadores o marcas de párrafo y éstos se incluyen como parte de los casos de estudio de las posibles alineaciones de las cuentas:

- Una oración en el texto O corresponde a una oración en el texto T(relación 1:1).
- Una oración en el texto O corresponde a dos oraciones en el texto T(relación
- 1:2).
- Dos oraciones en el texto O corresponden a una oración en el texto T(relación
- 2:1).
- Una oración en el texto O es eliminada en el texto T(relación 1:0).
- Una oración que no estaba en el texto O es agregada en el texto T(relación 0:1).
- Una marca de párrafo en inglés sin correspondencia.

- Una marca de párrafo en francés sin correspondencia.

- Una marca de párrafo en inglés corresponde a una marca de párrafo en francés.

La alineación a nivel de oraciones se lleva a cabo en dos etapas:
1. En la primer etapa se lleva a cabo una primera alineación con base en ciertas anclas en los textos O y T.
2. En la segunda etapa se lleva a acabo la alineación de las oraciones que se encuentran en las secciones en que dividen las anclas al corpus.

En la primer etapa de la alineación se elijen como anclas los comentarios que aparecen a lo largo del texto en donde se hace referencia a los anuncios realizados en el Parlamento (presidente, miembro, conjunto de miembros, inicio de receso, inicio de sesión, etc.). Esto se dividen en dos clases: anclas menores que abarcan a las intervenciones de los participantes y anclas mayores que se refiere al resto de comentarios que aparecen en el corpus.

La alineación de las anclas se realiza primero alineando las anclas mayores asignando un costo que favorece las verdaderas correspondencias y penaliza correspondencias confusas. A continuación se aceptan las secciones divididas por las anclas mayores si estas contiene la misma cantidad de anclas menores para ambos idiomas, en caso contrario se rechazan; a su vez las anclas menores realizan una subdivisión del corpus en subsecciones.

En la segunda etapa se utiliza un modelo del lenguaje que asume que las oraciones que aparecen en cada una las secciones en que se dividió el corpus fueron generadas por dos procesos aleatorios: el primero produce una secuencia de cuentas y el segundo determina las longitudes de las oraciones en cada cuenta. Ambos procesos forman un modelo de Markov para la generación de pares alineados.

Dada una cuenta se determina la longitud de las oraciones que la conforman, como sigue:

- La probabilidad de que la longitud de una oración en inglés l_e dado que no tiene correspondencia es igual a la probabilidad de obtener una oración de longitud l_e en el texto en inglés. De manera similar para el caso de una oración de longitud l_f en el texto en francés.

- La probabilidad de que una oración de longitud l_f corresponda con una oración de longitud l_e depende de la razón de las longitudes $r = log\,(l_f/l_e)$, que se asume esta normalmente distribuida con media μ y varianza s^2:

$$P(l_f|l_e) = \alpha exp\left[-(r-\mu)^2 / (2s^2)\right] \tag{3.6}$$

  donde α se elige de tal forma que la suma de $P\,(l_f|l_e)$ sea igual a la unidad.

- La probabilidad de dos oraciones en inglés con longitudes l_{e1} y l_{e2} corresponda a una oración en francés con longitud l_f esta dada por la razón de la longitud de la oración en francés y la suma de las longitudes de la oraciones en inglés que se encuentra normalmente distribuida con los mismo parámetros.

Variando los parámetros del modelo se obtiene que solamente el uso de marcas de párrafo conlleva a una tasa de error del 2%, mientras que solamente el uso de anclas reportó una tasa de error del $2{,}3\%$. El algoritmo presenta un buen desempeño para las alineaciones 1:1 y para algunos casos difíciles.

3.2.2. Metodologías estadísticas que usan información léxica

A diferencia de las metodologías anteriores, ahora se hace uso de información léxica para obtener pistas que ayuden en el proceso de alineación de las oraciones. El uso de estas características provee información útil que facilita la evaluación de la similitud entre oraciones, así estos métodos reportan una mayor precisión y menor vulnerabilidad al tipo de textos que se esta alineando, al mostrar tener mejores resultados para los casos en que existe sustracción y adición de oraciones, así como para los casos en que existe asignación múltiple. En términos generales se obtienen metodologías de alineación más robustas en comparación con las metodologías estadísticas.

Algunos de los trabajos de este tipo que tratan el problema de la alineación de textos paralelos a nivel de oraciones son Kay & Röscheisen [1993] y Chen [1993].

Además del uso de las propiedades léxicas, existen múltiples recursos léxicos disponibles que pueden utilizarse en la tarea de alineación como por ejemplo lista de cognados, analizadores morfológicos, diccionarios bilingües, entre otros.

De los primeros sistemas que utilizaron diccionarios bilingües tenemos el caso de Haruno & Yamazaki [1997] y Collier et al. [1998]. Estas metodologías se centran en el análisis de los idiomas inglés y los idiomas asiáticos como el japonés; utilizan generalmente dos tipos de diccionarios: uno permite relacionar los diferentes alfabetos que existen en el idioma japones, el segundo diccionario bilingüe establece una relación entre el idioma inglés y alguno de los alfabetos del japonés, principalmente para mostrar los límites equivalentes de una oración en ambos idiomas, ya que existen demasiadas diferencias estructurales entre ellos.

De especial interés es el trabajo presentado por Gautam & Sinha [2007], en este trabajo se desarrolla un sistema alinear dos textos paralelos Inglés-Hindi a nivel de oraciones utilizando un diccionario bilingüe y un algoritmo genético; el aspecto de interés es que se utiliza la información del diccionario de manera directa para para determinar si dos oraciones son candidatas a ser alineadas. A continuación se da una breve descripción del trabajo.

Gautam & Sinha (2007)

En este trabajo se describe un método de alineación que utiliza información léxica y estadística para determinar las cuentas que conforman la alineación. Utilizan una suma ponderada de un conjunto de parámetros estadísticos (que se obtienen de los textos paralelos que serán alineados) así como parámetros de correspondencia de acuerdo con la información de un diccionario bilingüe. El diccionario se considera que no está completo (no contiene un listado exhaustivo del léxico ni un listado completo de traducciones para cada una las entradas del diccionario).

Proponen la siguiente expresión para evaluar una cuenta candidata:

$$\begin{aligned} Eval\,() = {} & Complete\_Match * Exact\_match\_count \\ & + \sum_{todaslapalabras} \left[\frac{Word\_match}{Meaning\_count} \right] \\ & + \frac{Word\_factor}{\left| Word\_length\_ratio - \frac{dst\_words}{src\_words} \right|} \\ & + \frac{Char\_factor}{\left| Char\_length\_ratio - \frac{dst\_chars}{src\_chars} \right|} \end{aligned} \tag{3.7}$$

- *Meaning_count*:Número de significados que tiene una palabra en particular de acuerdo con el diccionario.
- *src_chars y dst_chars*:El número de caracteres en la oración origen y en la oración destino respectivamente.
- *src_words y dst_words*: El número de palabras en la oración origen y en la destino respectivamente.
- *Word_length_ratio*: La razón del número de palabras en el texto origen y el número de palabras en el texto destino.
- *Char_length_ratio*: La razón del número de caracteres en el texto origen y el número de caracteres en el texto destino respectivamente.

Se definen las siguientes variables de proporcionalidad que son optimizadas utilizando un algoritmo genético y una porción del texto a alinear:

- *Complete_match*
- *Word_match*
- *Word_factor*
- *Char_factor*
- *Threshold*: umbral que determina si dos oraciones son traducciones mutuas.
- *Window_percent*: ventana que indica la zona de búsqueda de la posible oración que se asignará en el texto destino.

La alineción de un texto se realiza en dos fases: en la primer fase se optimizan las variables mencionadas anteriormente utilizando un algoritmo genético simple y una porción del texto alineada manualmente; en la segunda fase se buscan las cuentas utilizando la expresión anterior para obtener como resultado la alineación del texto.

Se utilizó un corpus de prueba traducido de forma manual del hindi al inglés, que se asume esta alineado a nivel de párrafos. Los resultados obtenidos en este corpus se muestran el la tabla 3.3.

Tabla 3.3: Resultados obtenidos

Archivo	Precisión	Recall
Archivo de entrenamiento	82.94 %	88.43 %
Archivo de prueba 1	78.43 %	73.39 %
Archivo de prueba2	73.91 %	66 %
Archivo de prueba3	80.71 %	86.79 %

3.2.3. Metodologías léxicas

En este tipo de metodologías se hace uso principalmente de información léxica para realizar la alineación de los textos. Principalmente se utilizan como recursos diccionarios bilingües, tablas de correspondencia entre palabras, analizadores morfológicos, entre otros. El uso de recursos léxicos provee información confiable que facilita la evaluación de la similitud entre oraciones, así estos métodos reportan una efectividad considerable por lo que generalmente se aplican a textos que presentan una alta dificultad en dos sentidos: por el tipo de texto o porque los idiomas involucrados son completamente diferentes en su estructura. La principal desventaja que se tiene es que los métodos son más difíciles de implementar ya que requieren de recursos que no siempre se obtienen fácilmente, requieren de un procesamiento especial de los textos y se encuentran limitados a un conjunto de idiomas en especifico.

Gelbukh & Sidorov (2006)

Existen pocos trabajos al respecto, sin embargo, el trabajo presentado por Gelbukh & Sidorov [2006a] es de especial interés ya que en este trabajo se presenta una función de evaluación, la cual sirvió de base para el método desarrollado en el presente trabajo, que utiliza traducciones mutuas para determinar la similitud entre párrafos de textos literarios.

Se propone la función $Dictionary\ (So,\ ST)$ que regresa el número de palabras significativas (tokens) que no son traducciones mutuas entre dos textos. La función $Dictionary\ (So,\ ST)$ se muestra en la ecuación 3.8

$$Distance\,(T_A, T_B) = |T_A| + |T_B| - 2 * traslations \qquad (3.8)$$

Donde:

- T_{A1}: representa el número de *tokens* que contiene un párrafo en especifico del texto en el idioma A.
- T_{B1}: representa el número de *tokens* que contiene un párrafo en especifico del texto en el idioma B.
- *translations* corresponde al número de traducciones mutuas que contienen T_{A1} y T_{B1}. Se considera que dos *tokens* son traducciones mutuas si se cumplen las siguientes condiciones:
    1. Si los dos *tokens* son traducciones mutuas de acuerdo con un diccionario bilingüe (como *word types*).
    2. Los párrafos en que ocurren se suponen alineados.

El método utiliza una representación basada en gráfos y un algoritmo de programación dinámica para encontrar la mejor alineación posible entre un par de textos paralelos bilingües español–inglés. El método reporta una efectividad superior al 90% para diferentes tipos de alineaciones(1:1, 1:2, 1:3, 2:1, 3:1) en un corpus de textos literarios.

Capítulo 4

Método propuesto

4.1. Introducción

En este trabajo de tesis, se considera a la tarea de alineación de corpus paralelo a nivel de oraciones como un problema de optimización que consiste en buscar la alineación particular del corpus paralelo que sea la que obtenga la mejor evaluación con base en un modelo propuesto que utiliza información léxica y estadística de los idiomas español e inglés. La búsqueda de la mejor alineación se resuelve utilizando un enfoque de programación dinámica.

Una alineación candidata o simplemente una alineación a nivel de oraciones, denotada por A_i, corresponde a una configuración particular de oraciones alineadas, es decir, oraciones que se asume son traducciones mutuas. La alineación con la mejor evaluación para un corpus paralelo español–inglés en particular, se llamará mejor alineación o solución y será denotada por A_{best}.

El método propuesto realiza la búsqueda de la mejor alineación (A_{best}) en términos de oraciones alineadas, encontrando la mejor configuración para cada una de las oraciones contenidas en el corpus paralelo. La elección de la mejor configuración para cada oración se realiza considerando la conservación del contenido entre las oraciones de los distintos idiomas y el orden de aparición dentro del texto al que pertenecen.

Debido a que el método propuesto se enfoca en la problemática presentada en corpus paralelo con traducción literaria, éste no solamente se centra en los tipos de alineaciones comunes (1:1, 1:2, 2:1), incluye también los casos de omisión, inserción y alineación múltiple. De manera general el método resuelve alineaciones de la forma m:n donde $m \in \{\vee, 1, 2, \ldots\}$ y $n \in \{\vee, 1, 2, \ldots\}$,

exceptuando el caso en que ambos m y n sean al mismo tiempo el elemento vacío ($\vee$).

En las secciones siguientes se describe con más detalle las características del método propuesto.

4.2. Modelo propuesto

4.2.1. Representación

Una unidad de correspondencia B se utiliza para designar a dos subconjuntos de oraciones (del texto en el idioma origen y del texto en el idioma destino respectivamente) que son traducciones mutuas (también llamadas alineadas).

Definimos una unidad de correspondencia como sigue:

Definición 4.1 *Sea T_O el conjunto que contiene a todas las oraciones del texto en el idioma origen y T_T el conjunto de todas las oraciones del texto en el idioma destino. Una unidad de correspondencia B es una tupla (S_O, S_T) donde:*

- $S_O \subset T_O$ *es un subconjunto ordenado de oraciones del texto en el idioma origen.*
- $S_T \subset T_T$ *es un subconjunto ordenado de oraciones del texto en el idioma destino.*

En una unidad de correspondencia B se dice que las oraciones en S_T son traducciones de las oraciones en S_O, o bien, que las oraciones en S_O son asignadas a las oraciones en S_T y las oraciones se encuentran ordenadas de acuerdo con el orden de aparición en el texto. Se utilizará el conjunto vacío ($\vee$)

para decir que no existe correspondencia en el texto, principalmente para los casos de eliminación e inserción de oraciones (véase sección 2.1.2), salvo el caso particular en que S_O y S_T son vacíos al mismo tiempo, el cual será ignorado.

Una alineación a nivel de oraciones para un par de textos paralelos o simplemente una alineación, es el listado de las oraciones del texto en el idioma origen con sus respectivas traducciones del texto en el idioma destino. Podemos definir a una alineación de la siguiente manera:

Definición 4.1 *Una alineación A es un conjunto ordenado de unidades de correspondencia B*

$$A = \{B_1, \ldots, B_k\}$$

tal que:

- *Toda oración del texto en idioma origen y del texto en el idioma destino debe ser alineada, esto es,*

$$\forall o_i \in T_O, \exists B_j \; tal \; que \; o_i \in S_{Oj}$$

y

$$\forall t_i \in T_T, \exists B_m \; tal \; que \; t_i \in S_{Tm}$$

- *No deben existir referencias cruzadas entre las oraciones de dos cuentas distintas, es decir, si la i-ésima oración del texto en el idioma origen o_i fue asignada a la oración t_k del texto en el idioma destino, entonces la oración o_{i+1} no puede ser asignada a una oración anterior a t_k.*

Una alineación en particular depende de las unidades de correspondencia que la conforman, las cuales generalmente son de alguno de los tipos presentados en la sección 2.1.2.

En la figura se muestra un ejemplo de representación de alineaciones para el texto paralelo que se muestra en la tabla 4.1.

Tabla 4.1: Ejemplo de un corpus de prueba

Texto en inglés		Texto en español	
No.	Texto	No.	Texto
1	Mexico is a beatiful country	1	México es un país bonito
2	Mexico´s beaches are beatiful	2	Las playas de México son muy bonitas
3	The tacos are a typical dish from Mexico	3	Los tacos son un platillo típico de México

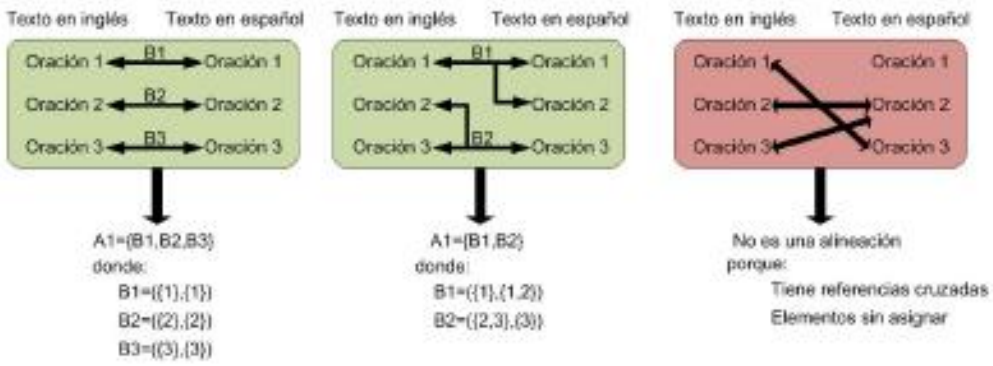

Figura 4.1: *Ejemplos de alineaciones para el texto de la tabla 4.1*

4.2.2. Función de evaluación

La calidad de una alineación en particular será determinado por una función de evaluación $Score()$, la cual asignará un valor a una alineación A_i para un par de textos paralelos T_O y T_T.

Como se había mencionado anteriormente, una alineación A_i se compone de una secuencia de unidades de correspondencia $(B_1, \ldots, B_k)$. Una función llamada $Similitud()$ asigna una puntuación a cada unidad de correspondencia B_i y la suma de cada una de las puntuaciones de las unidades de correspondencia es igual a la puntuación de la alineación, esto es,

$$Score(A_i, T_O, T_T) = \sum_{j=1}^{k} Similitud(B_j) \qquad (4.1)$$

En la figura 4.2 se ilustra la forma de evaluar una alineación.

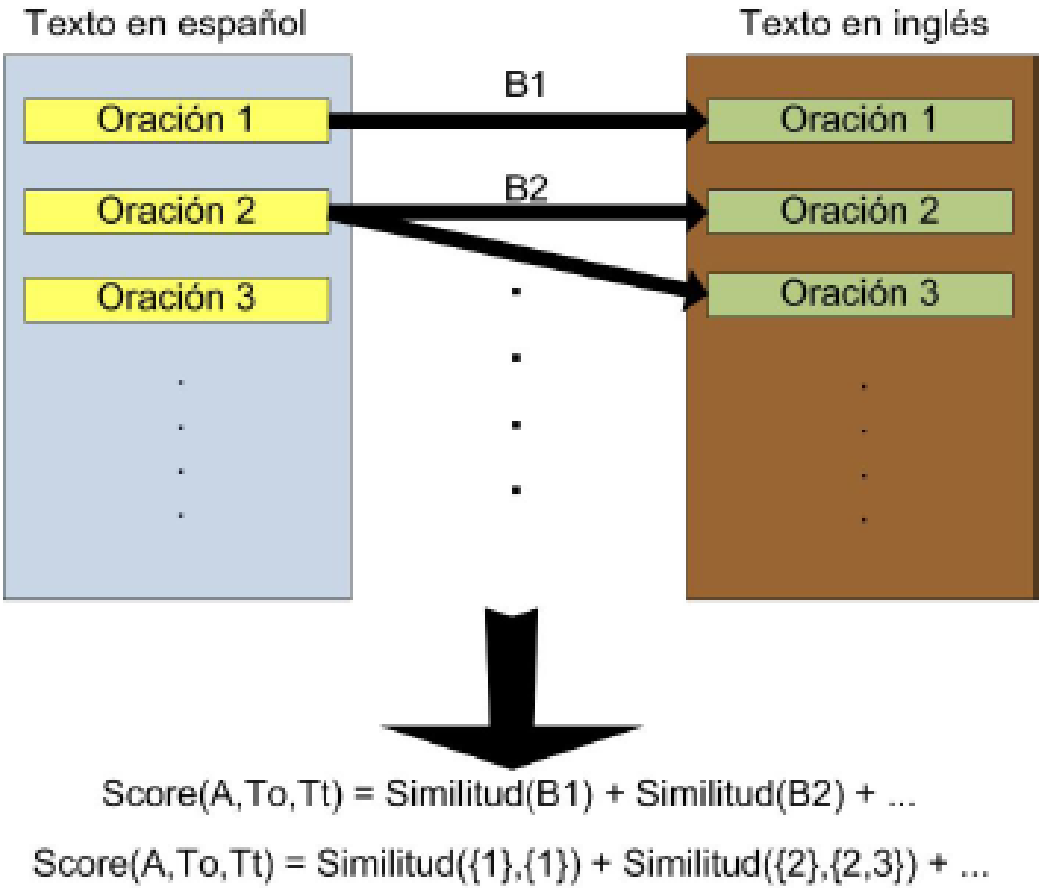

Figura 4.2: *Diagrama que ilustra la forma de evaluar una alineación*

Una unidad de correspondencia B_i puede tener distintas configuraciones, esto depende de las oraciones que la conforman, es decir, depende de la elección de S_O y de S_T.

La función $Similitud()$ que se propone para evaluar cada una de las unidades de correspondencia toma en cuenta tres características que en diversos trabajos han demostrado tener buenos resultados en alineación de corpus paralelo, ya que son características que se encuentran altamente correlacionadas tanto en el idioma inglés como en el idioma español: traducciones mutuas [Gelbukh & Sidorov 2006a], número de palabras [Brown et al. 1991b] y número de caracteres en la oración [Gale & Church 1991].

La función $Similitud()$ que se propone se muestra en la ecuación 4.2.

$$Similitud\,(S_O, S_T) = Dictionary\,(S_O, S_T) + MeanLength\,(S_O, S_T) + CharLength\,(S_O, S_T) \tag{4.2}$$

La función $MeanLength(S_O, S_T)$ regresa el valor absoluto de la diferencia entre el número de ***elementos significativos*** presentes en S_O y el número presente en S_T.

La función $CharLength(S_O, S_T)$ regresa el valor absoluto de la diferencia entre el número de caracteres presentes en S_O y el número presente en S_T.

La función $Dictionary\,(S_O, S_T)$ se propone en el trabajo Gelbukh & Sidorov [2006a], esta función regresa el número de palabras significativas que no son *traducciones mutuas* entre S_O y S_T. En este trabajo proponemos utilizar una extensión de la idea original utilizando el concepto de ***elementos significativos*** el cuál se refiere a aquellas aquellos elementos que son representativos de una oración, es decir, consideramos como significativos a los siguientes elementos: ***sustantivos***, ***adjetivos***, ***verbos***, ***adverbios***, ***abreviaturas***, ***números***, ***nombres propios***, ***signos de admiración e interrogación***.

El término ***traducciones mutuas*** se refiere a elementos en el idioma origen cuyos lemas ***aparezcan como traducciones*** de algún elemento en el idioma destino de acuerdo con un diccionario bilingüe.

La función $Dictionary\,(S_O, S_T)$ se muestra en la ecuación 4.3

$$Dictionary\,(S_O, S_T) = MElements\,(S_O) + MElements\,(S_T) - 2 * Trad\,(S_O, S_T) \tag{4.3}$$

Donde:

- $MElements()$: es una función que toma como argumento una oración en algún idioma y regresa el número de **elementos significativos** que contiene.
- $Trad(S_O, S_T)$: es una función que toma como argumento dos conjuntos de oraciones S_O y S_T y regresa el número de **elementos significativos** que son **traducciones mutuas**.

En el modelo presentado, las unidades de correspondencia serán las variables que se modificaran en busca de la mejor alineación y las características del corpus paralelo serán los parámetros que se utilizarán en la función de evaluación (ecuación 4.1 y ecuación 4.2), que por la forma en que éstas fueron definidas, $\mathcal{A}_{best}$ será aquella alineación que reciba la menor evaluación.

4.3. Implementación del método

En la sección 4.2 se ha descrito el modelo utilizado en la propuesta de solución, en esta propuesta se utilizan dos características de ámbito lingüístico (traducciones mutuas y elementos significativas), así como, el número de caracteres que conforman a cada oración.

Teniendo en cuenta las características de la propuesta de solución, se implementó un sistema para la alineación de un corpus paralelo, en el que se tiene como entrada un corpus paralelo español–inglés y se obtiene como salida la mejor alineación del corpus en un listado donde se indica la correspondencia de las oraciones del corpus. En la figura 4.3 se muestra el diagrama a bloques del sistema, cada una de las etapas se describen a continuación.

4.3.1. Etapa de preprocesamiento

En esta etapa se realizan principalmente dos tareas: el filtrado de caracteres y la segmentación del texto en oraciones. Se asume que el corpus de entrada corresponde a un texto plano sin formato y con una estructura simple (el texto no contiene ninguno de los siguientes elementos: citas cruzadas, notas al pie de página, imágenes, tablas, encabezados, viñetas, ecuaciones, citas bibliográficas, índices y glosarios).

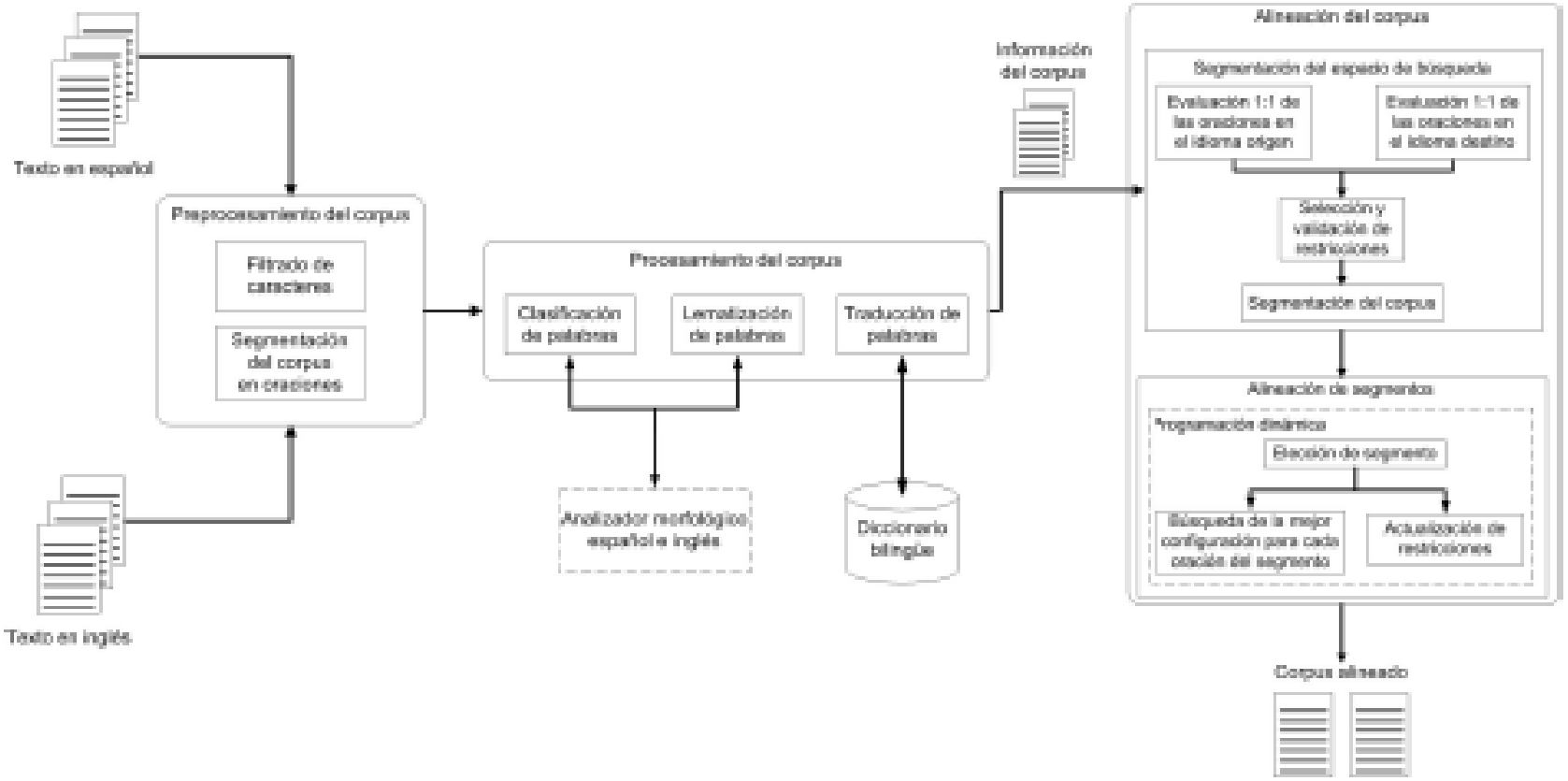

Figura 4.3: *Diagrama a bloques de la propuesta para la alineación a nivel de oraciones*

La tarea de filtrado de caracteres se realiza en principio eliminando del corpus aquellos caracteres que no serán utilizados con el fin de simplificar la etapa de segmentación del texto; para ello se propone eliminar aquellos caracteres que no se encuentran en el alfabeto que se describe a continuación:

- Las letras mayúsculas A, B, C, . . . , Z y las letras minúsculas a, b, c, . . . , z.
- Números (0, 1, . . . , 9);
- Signos de puntuación: punto (.), coma (,), punto y coma (;), dos puntos (:), puntos suspensivos (. . .), signo de admiración (¡ !), signos de interrogación (¿ ?) y apóstrofo.
- Salto de línea.
- Espacio.
- caracteres especiales: $ y &

La segmentación del corpus paralelo en oraciones es la tarea más importante que se lleva a cabo en esta etapa, porque en ella se identifican las oraciones presentes en el texto, es decir, las unidades con las que se va a trabajar en las etapas siguientes. En esta etapa no se incluye algún otro indicador de división del texto en un orden superior como párrafos, secciones o capítulos.

La segmentación de un texto en oraciones depende en gran medida de la estructura del idioma con la que se trabaja porque en algunos casos no existen signos de puntuación que funcionen como delimitadores o por el contrario se cuenta con varios signos que funcionan como delimitadores. Algunos ejemplos de trabajos realizados al respecto son: Kim et al. [2000], Xiong et al. [2009] y Li et al. [1990].

Una oración se define como[1]:
“Palabra o conjunto de palabras que se expresan en un sentido gramatical completo”.

En el caso particular de los idiomas inglés y español típicamente una oración se encuentra delimitada por un punto (.) y un espacio, sin embargo, existen algunos otros casos como por ejemplo cuando se utiliza un punto y coma (;), en los que se

[1] 1Véase Diccionario de la Real Academia de la Lengua Española, Junio 2011, http://buscon.rae.es/draeI/SrvltConsulta?TIPO_BUS=3&LEMA=oración.

requiere realizar un análisis gramatical para saber si se utiliza como un delimitador de oraciones o no.

En este trabajo no se realiza una segmentación exhaustiva del corpus, simplemente se utilizan reglas basadas en los signos de puntuación para efectuar la segmentación del corpus en oraciones. Las reglas que se utilizan son[2]:

1. Una oración es una secuencia de palabras que termina en alguno de los siguientes casos:
   - Termina con un punto (.).
   - Termina con un salto de línea.
   - Termina con un signo de admiración (!) y la palabra siguiente comienza con una letra mayúscula o se presenta un signo de interrogación (¿) o se presenta un signo de admiración (¡).
   - Signo de interrogación (?) y la palabra siguiente comienza con una letra mayúscula o se presenta un signo de interrogación (¿) o se presenta un signo de admiración (¡).
   - Dos puntos (:) y la palabra siguiente comienza con una letra mayúscula o se presenta un signo de interrogación (¿) o se presenta un signo de admiración (¡).
   - Puntos suspensivos(. . .) y la palabra siguiente comienza con una letra mayúscula o se presenta un signo de interrogación (¿) o se presenta un signo de admiración (¡).
2. Los títulos se consideran oraciones.
3. Números utilizados en una enumeración se consideran oraciones.
4. Los elementos de una enumeración se consideran oraciones.

4.3.2. Etapa de procesamiento del corpus

En esta etapa se lleva a cabo un procesamiento de los elementos que conforman a las oraciones (palabras, signos de puntuación y números); lo que se busca con dicho procesamiento es: identificar aquellas elementos que son representativos de una oración y obtener la traducción de dichos elementos. Al final de esta etapa se obtiene la información necesaria para aplicar el algoritmo de programación dinámica y buscar la mejor alineación.

En esta etapa se llevan a cabo tres tareas:

La primer tarea consiste en clasificar a los elementos para identificar aquellos que serán de utilidad.

---

[2] [2]Véase The BAF: A Corpus of English–French Bitext,Michael Simard, Junio 2011, http://www.iro.umontreal.ca/˜simardm/lrec98/.

La segunda tarea consiste en obtener una representación estándar de los elementos (en el caso particular de las palabras se hace obteniendo sus lemas).

La tercer tarea consiste en obtener la traducción de cada uno de los elementos (en el caso particular de las palabras se hace traduciendo los lemas mediante un diccionario bilingüe).

Cuando se habla de traducción el punto central es la preservación del significado. En una oración la mayor cantidad de información recae en ciertas palabras a las cuales denominamos palabras de contenido; para un gran número de idiomas las palabras de contenido generalmente pertenecen a las siguientes categorías gramaticales: sustantivos, verbos, adjetivos y adverbios [Papageorgiou et al. 1994].

Un elemento significativo es aquella palabra o elemento que es representativo de una oración. Los elementos significativas incluyen a las palabras de contenido, nombres propios, números, abreviaturas, signos de admiración y signos de interrogación.

Para resolver la primer tarea se deben identificar a los elementos significativos de una oración, por lo que se clasifica a los elementos de acuerdo con las siguientes clases:

1. Abreviatura: se refieren a aquella palabra que comienza por una letra mayúscula, termina con un punto y tienen una longitud menor a 6 caracteres.
2. Signo de puntuación: solamente se toman en cuenta los signos de cierre de interrogación (?) y de cierre de admiración (!).
3. Número: cualquier cadena que contenga números y puntos.
4. Palabra de contenido: se refiere a sustantivos, adjetivos, verbos y adverbios.
5. Palabra auxiliar: se refiere a los artículos, preposiciones, conjunciones, pronombres e interjecciones.
6. Nombre propio: se refiere a palabras que comienzan con una letra en mayúscula, tienen solamente un lema y solo una traducción para ese lema.

Para clasificar a un elemento dentro de las clases 1,2 y 3 se desarrollaron heurísticas basadas en la ocurrencia de ciertos patrones de caracteres, mientras que para las clases 4, 5 y 6 se emplearon analizadores morfológicos. Para el caso de las palabras en español se utilizó el sistema AGME [Sidorov et al. 2002] que tiene un diccionario morfológico de 26,000 entradas que equivale a más de 1,000,000 de formas gramaticales; para el caso de las palabras en inglés se utilizó un analizador morfológico [Gelbukh & Sidorov 2006b] basado en un diccionario morfológico de WordNet de alrededor de 60,000 entradas.

Al mismo tiempo que los elementos son clasificados, se resuelve la segunda tarea que consiste en expresar a los elementos en una forma estándar que facilite su manipulación posterior. En el caso particular de las palabras se obtienen todos los lemas de las palabras utilizando los analizadores morfológicos, mientras que para

el resto de las clases los elementos permanecen como aparecen en el texto original (números, abreviaturas y signos de puntuación).

Para la tercer tarea los elementos que se encuentren en la categoría de palabras auxiliares no serán traducidos y se descartan al momento de evaluar la correspondencia entre oraciones debido a que estas palabras son de uso muy frecuente en los idiomas español e inglés y pueden causar confusión al asignar la puntuación de similitud. Para las demás clases a excepción de las palabras de contenido se traducen tal como aparecen en el texto original ya que serán considerados como cognates y finalmente para el caso de las palabras de contenido se traducen cada uno de sus lemas utilizando el diccionario bilingüe inglés–español VOX que se encuentra disponible en la Web.

Con la información obtenida en esta etapa se generan dos archivos(uno por cada texto) de metadatos codificados en XML donde se expresa la división del texto en oraciones, la división de las oraciones en elementos, la clase a la que pertenece cada elemento, los lemas en caso de que el elemento sea una palabra de contenido, las traducciones del elemento si existen en el diccionario e información adicional sobre la oración como: número de elementos, número de elementos significativos y número de caracteres. En la figura 4.4 se muestra la forma en que se estructura la información en los archivos y en la figura 4.5 se muestra un ejemplo de un archivo generado con el corpus de prueba.

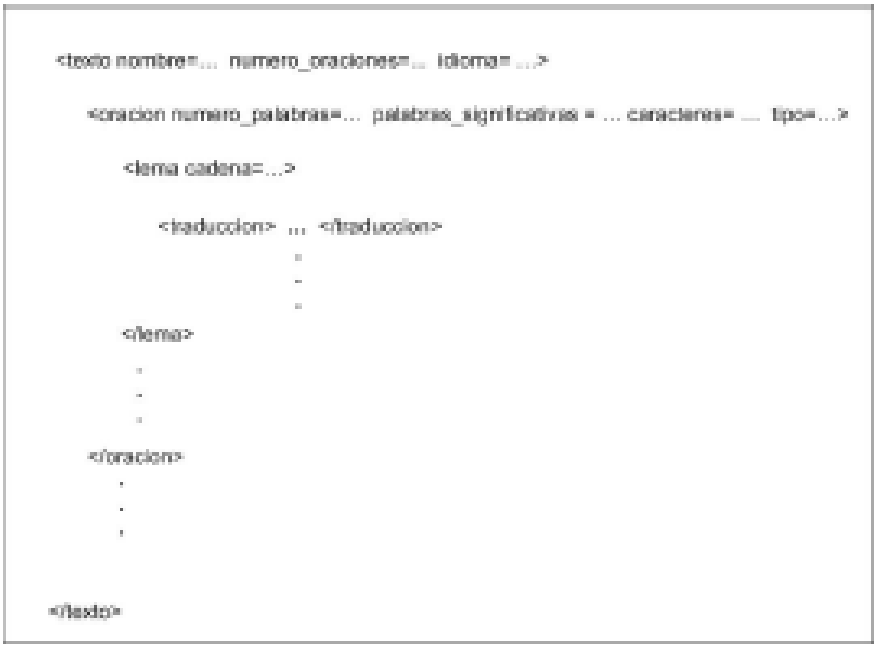

Figura 4.4: *Esquema del archivo XML que contiene la información del texto*

Figura 4.5: *Ejemplo de un archivo XML que contiene la información de un texto*

4.3.3. Etapa de alineación del corpus

En esta etapa se busca la mejor alineación A_{best} del corpus paralelo. De acuerdo con la definición 4.1 el problema se puede resolver encontrando las cuentas que presenten la mejor puntuación de acuerdo con la ecuación 4.2 y que además se satisfagan las siguientes condiciones:

- Toda oración del corpus debe ser alineada.
- No deben presentar referencias cruzadas.

Por la forma en que se ha definido el modelo a utilizar, queda claro que la elección de una unidad de correspondencia B_i depende de las unidades de correspondencia elegidas anteriormente $(B_{i-1}, \ldots, B_1)$, palabras tienen una frecuencia de ocurrencia baja, por lo tanto, utilizando la correspondencia de elementos significativos entre oraciones de distintos idiomas es posible encontrar pares de oraciones que puedan utilizarse para simplificar el espacio de búsqueda bajo el supuesto de que no existen referencias cruzadas.

El algoritmo de búsqueda que se propone se compone de dos fases:

- En la primer fase se realiza un búsqueda exhaustiva de restricciones adicionales (también conocidos como puntos ancla) sobre el corpus de la forma 1:1, que ayuden a segmentar el espacio de búsqueda en subespacios más pequeños.
- En la segunda fase se utiliza un algoritmo de programación dinámica para realizar la búsqueda de la mejor alineación (A_{best}) a través de los nuevos subespacios de búsqueda.

Primer fase

La primer fase del algoritmo de búsqueda se basa en el siguiente hecho: en una alineación independientemente del tipo de texto que se utilice, existe la tendencia de que la mayor parte de las unidades de correspondencia que la conforman sean del tipo 1:1.

Lo que se propone en esta primer fase es realizar una búsqueda de nuevas restricciones bajo la forma de oraciones alineadas del tipo 1:1 que presenten una alta tendencia a pertenecer a la misma unidad de correspondencia.

En esta fase se realiza la una búsqueda exhaustiva por cada una de las oraciones del texto en el idioma origen en las oraciones del texto en el idioma destino. Como función de evaluación se utiliza la ecuación 4.3 y se eligen aquellos pares que presente la menor evaluación.

Una vez que se ha elegido a la pareja de cada una de las oraciones en el idioma origen, se realiza la misma búsqueda exhaustiva pero ahora para la oración en el idioma destino seleccionada.

A continuación se comparan las tuplas seleccionadas elemento a elemento, si éstas coinciden completamente se almacenan temporalmente para el siguiente paso, en caso contrario se descartan. Finalmente se realiza un última verificación sobre los pares seleccionadas eliminando aquellos pares que ocasionen referencias cruzadas entre ellos.

Al termino de la primer fase del algoritmo se obtiene un listado de pares de oraciones con una alta probabilidad de pertenecer a una misma unidad de correspondencia de manera completa o parcial verificado con la información contenida en el diccionario bilingüe.

En la primer fase del algoritmo de búsqueda tenemos los siguientes elementos:

- ENTRADA: Dos archivos de metadatos que contengan la información del corpus paralelo a alinear, uno por cada idioma (véase sección 4.3.2).
- PROCESO: Un par de oraciones se considera una restricción si para cada elemento su pareja asociada presenta la mejor evaluación de acuerdo con la ecuación 4.3 y no genera una referencia cruzada con las demás restricciones.
- SALIDA: Un archivo que contiene el listado de pares de oraciones que se consideran restricciones.

Los pasos que se realizan en la primer fase del algoritmo son:

1. Seleccionar una oración en el texto del idioma origen.
2. Evaluar la oración seleccionada con cada una de la oraciones en el texto del idioma destino utilizando la ecuación 4.3.
3. Almacenar aquella oración con la que obtuvo la mejor evaluación.
4. Evaluar la oración seleccionada con cada una de las oraciones en el texto del idioma destino.
5. Seleccionar aquella oración con la que obtuvo la mejor evaluación.
6. Comparar las elecciones realizadas por las oraciones, si coinciden almacenar como restricción tentativa, en caso contrario descartar.
7. Seleccionar una nueva oración del texto en el idioma origen.
8. Repetir los pasos del 2 al 6 hasta que se hayan evaluado todas las oraciones del texto en el idioma origen.
9. Eliminar aquellas restricciones que causen referencias cruzadas.

En la figura 4.6 se muestra el diagrama de la primer fase del algoritmo de búsqueda y a continuación se da una explicación de las variables:

- O_i: Variable de tipo entero que indica el número de oración del texto origen que se esta analizando.
- O_best: Variable de tipo entero que indica el número de oración en el texto destino que representa la mejor pareja para la oración O_i.
- T_best: Variable de tipo entero que indica el número de oración en el texto origen que representa la mejor pareja para la oración O_best.

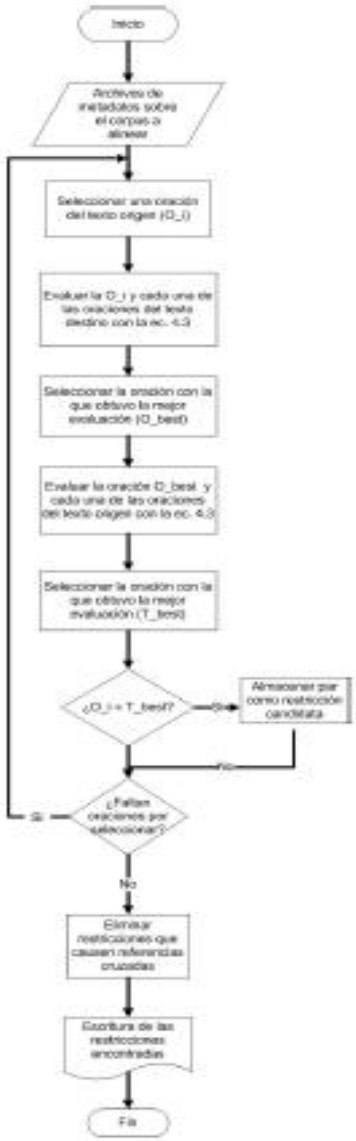

Figura 4.6: *Diagrama de flujo para la primer fase de la alineación*

Segunda fase

En esta segunda fase se realiza propiamente la búsqueda de la mejor alineación (A_{best}) utilizando un enfoque de programación dinámica. Basándose en este enfoque la búsqueda se puede realizar dividiendo el problema original en subproblemas en los que se desea encontrar la mejor configuración para cada una de las oraciones que conforman el corpus que se desea alinear.

En la segunda fase del algoritmo se utilizan las restricciones obtenidas en la primera fase para realizar segmentaciones del espacio de búsqueda que ayuden a encontrar la mejor alineación A_{best}. Utilizando las restricciones encontradas podemos plantear la búsqueda de A_{best} como encontrar la mejor configuración posibles para cada una de las oraciones que se encuentren dentro de un segmento del espacio de búsqueda acotados por dos restricciones utilizando el modelo planteado en la sección 4.2.

La búsqueda de la mejor configuración para las oraciones contenidas en un segmento del espacio de búsqueda se puede realizar dividiendo la configuración de las oraciones en unidades de correspondencia, donde, la elección de una unidad de correspondencia depende de las unidades seleccionadas anteriormente y tomando en cuenta la restricción de que no existen referencias cruzadas.

Tomando en cuenta lo mencionado anteriormente, la búsqueda de las unidades de correspondencia se puede realizar de manera ordenada siguiendo el orden de aparición de las oraciones en el corpus.

Para una oración en particular contenida en determinado segmento, se puede representar el espacio de búsqueda de las unidades de correspondencia mediante un árbol en el que cada hoja representa un estado o unidad de correspondencia en particular del cuál derivan los siguientes estados posibles, tomando en cuenta las restricciones de la definición 4.1, donde la mejor configuración corresponde al recorrido desde el nivel superior del árbol hasta un nivel inferior que presente la mínima evaluación de acuerdo con la ecuación 4.2.

En la figura 4.7 se muestra un ejemplo de representación para un conjunto de oraciones contenidas en cierto segmento del espacio de búsqueda. Cada cuadro de la representación representa una unidad de correspondencia, en el que se encuentran inscritos a la derecha del guión las oraciones del texto en el idioma origen (en este caso de la oración 1 a la n) y a la izquierda del guión las oraciones del texto en el idioma destino (de la oración i a la i+k); los cuadros que se encuentran en un mismo nivel corresponden a los estados posibles que se pueden generar al haber elegido un estado previamente.

La mejor configuración posible para las oraciones contenidas en un segmento determinado, corresponde al camino elegido que posea la mejor evaluación de acuerdo con la función de evaluación.

En la segunda fase del algoritmo de búsqueda tenemos los siguientes elementos:

- ENTRADA: Un archivo que contiene las restricciones obtenidas en la primer fase. sección 4.3.2).
- PROCESO: La mejor alineación a nivel de oraciones para un corpus paralelo específico corresponde a las unidades de correspondencia con la mejor evaluación de acuerdo con la ecuación 4.2.
- SALIDA: Un archivo que contiene el listado de los identificadores de las oraciones que conforman las unidades de correspondencia de la mejor alineación (A_{best}).

Los pasos que se llevan a cabo en la segunda fase son:

1. Elegir un segmento del espacio de búsqueda.
2. Elegir una oración del texto en el idioma origen O_i.
3. Realizar la expansión de estados(**unidades de correspondencia**) que incluyan a la oración elegida y que se encuentre en el espacio factible del segmento.
4. Evaluar cada uno de los estados utilizando la ecuación 4.2.
5. Selecciona el mejor estado Uo_best.
6. Almacena el estado ganador Uo_best.

7. Elegir una oración del texto en el idioma destino T_i.

8. Realizar la expansión de estados(**unidades de correspondencia**) que incluyan a la oración elegida y que se encuentre en el espacio factible del segmento.

9. Evaluar cada uno de los estados utilizando la ecuación 4.2.

10. Selecciona el mejor estado Ut_best.

11. Almacena el estado ganador Ut_best.

12. Comparar la aptitud de los estados Uo_best y Ut_best.

13. Seleccionar al estado con la mejor aptitud.

14. Asignar la unidad de correspondencia ganadora U_best.

15. Actualizar las restricciones en el segmento.

16. Seleccionar una nueva oración del texto en el idioma origen.

17. Repetir los pasos del 3 al 16 hasta que se hayan evaluado todas las oraciones del texto en el idioma origen y del texto en el idioma destino.

18. Asignar las oraciones del corpus que no encuentren en ningún estado ganador al elemento *vacío*.

19. Se repiten los pasos del 1 al 18 hasta que se hayan seleccionado todos los segmentos en que se dividió el corpus.

En la figura 4.8 se muestra el diagrama de la segunda fase de alineación y a continuación se da una explicación de las variables:

- O_i: Variable de tipo entero que indica el número de oración del texto origen que se esta analizando.
- T_i: Variable de tipo entero que indica el número de oración del texto destino que se esta analizando.
- Uo_best: Variable que indica el estado con la mejor evaluación para la oración O_i.
- Ut_best: Variable que indica el estado con la mejor evaluación para la oración T_i.
- U_best: Variable que indica el estado con la mejor evaluación entre Uo_best y Ut_best número.

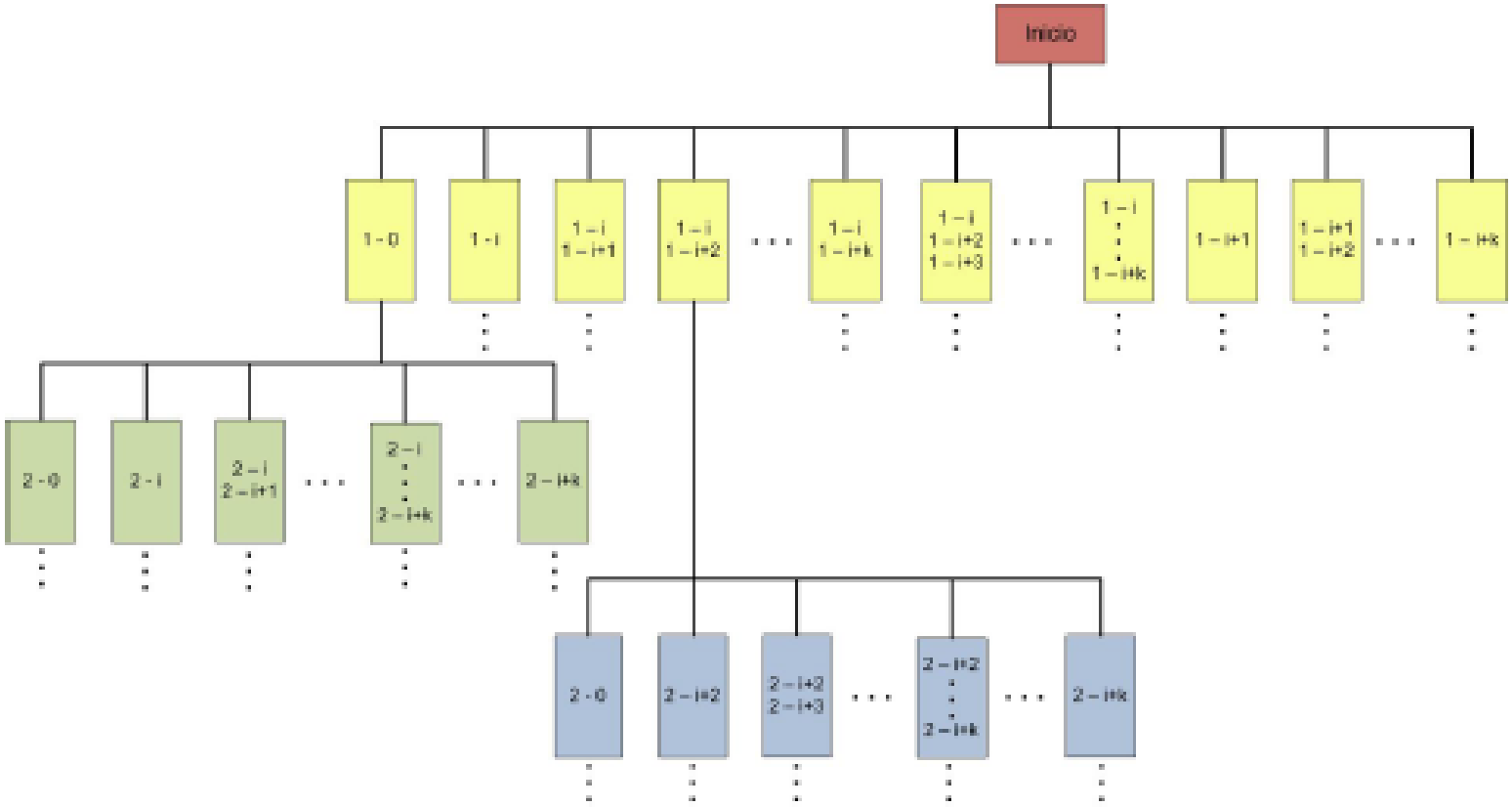

Figura 4.7: *Diagrama de árbol para el espacio de factible de un segmento del corpus*

Figura 4.8: *Diagrama de flujo para la segunda fase de la alineación*

Complejidad del algoritmo

La complejidad del algoritmo de búsqueda desde el punto de vista del número de veces que se utiliza la función de evaluación es de orden O_3, esto corresponde a la segunda fase del algoritmo bajo el supuesto de que el número de oraciones del texto en el idioma origen (S_O) sea aproximadamente igual al número de oraciones del texto en el idioma destino (S_T, esto es, $S_O \vee S_T$ y que el número de estados promedio por cada una de las oraciones sea S_O/k con $k \in N$. Para la primer fase del algoritmo la complejidad es de orden O_2 bajo el supuesto de que $S_O \vee S_T$.

Si bien el objetivo de la tesis nos se centra en la eficiencia del algoritmo de algoritmo de alineación sino en la efectividad de éste, se puede plantear como trabajo futuro realizar algunas adecuaciones al algoritmo para reducir la complejidad del mismo, por ejemplo, en la primer fase del algoritmo se puede realizar la búsqueda de las restricciones solo en los vecinos más cercanos de acuerdo con la posición relativa de las oraciones, en la segunda fase puede reducirse el número de estados que se obtienen para una unidad de correspondencia en particular al restringir el tipo de alineaciones posibles (véase sección 2.1.2).

Capítulo 5
Pruebas y resultados

Para la evaluación del método se utilizó como corpus de prueba algunos capítulos de textos literarios (específicamente novelas de ciencia ficción), escritos en los idiomas español e inglés, seleccionados de una colección recopilada por Gelbukh & Sidorov [2006a].

La prueba consiste en seleccionar un par de textos del corpus de prueba, posteriormente se alinea el par de textos utilizando el método propuesto y finalmente se comparan las unidades de correspondencia de cada una de las alineaciones con la alineación realizada por un humano del mismo texto para evaluar cada uno de los métodos.

En este caso se trata a cada unidad de correspondencia como un elemento indivisible al momento de comparar.

5.1. Corpus de prueba

Se realizaron experimentos utilizando como corpus de prueba algunos capítulos de novelas de ciencia ficción en los idiomas inglés y español. En la tabla 5.1 se muestra la información del corpus de prueba.

Se realizó la alineación manual(gold align) del corpus de prueba, por una persona con habilidad de lectura de textos en inglés y español. Para la obtención del gold align de cada par de textos paralelos que conforman el corpus de prueba, el humano fue provisto por una versión del corpus de prueba segmentada en oraciones por la implementación del método propuesto (véase sección 4.3.1).

En las tablas 5.2 y 5.3 se muestran la distribución de los tipos de alineación que presentan los textos de acuerdo con la alineación manual.

Tabla 5.1: *Corpus de prueba*

Autor	Título	Extensión
Conan Doyle, Arthur	*Las aventuras de Sherlock Holmes*	4 capítulos
James, Henry	*Otra vuelta de tuerca*	2 capítulos

Tabla 5.2: características del texto *Otra vuelta a la tuerca*

Sección	Tamaño (inglés)	Tamaño (español)	Tipo de alineación	No. de ocurrencias
capítulos 1–2	361 oraciones	363 oraciones	1:1	300
			1:2	12
			2:1	8
			2:2	1
			1:∅	1
			∅:1	1

Tabla 5.3: características del texto *Las aventuras de Sherlock Holmes*

Sección	Tamaño (inglés)	Tamaño (español)	Tipo de alineación	No. de ocurrencias
capítulos 1–4	2258 oraciones	2267 oraciones	1:1	2061
			1:2	65
			2:1	64
			2:2	4
			3:1	1
			1:3	2
			1:∅	11
			∅:1	7

5.2. Medida de evaluación

5.2.1. Efectividad del método

La **efectividad** (P) se define como la razón del número de ***unidades de correspondencia*** **correctas** entre el número **total de** ***unidades de correspondencia*** que arroja el sistema.

$$P = \frac{unidades\ de\ correspondencia\ correctas}{total\ de\ unidades\ de\ correspondencia\ obtenidas} \tag{5.1}$$

Se considera a una ***unidad de correspondencia correcta*** como aquella que coincide elemento a elemento con alguna unidad expresada en la alineación manual (*gold standar*).

La **efectividad para un tipo de alineación** i denotada como (P_i) se define como la razón del número de ***unidades de correspondencia*** **del tipo** *i* **correctas** que arroja el método entre el número **total de** ***unidades de correspondencia*** del tipo *i* que existen el texto (*gold standar*).

$$P_i = \frac{unidades\ de\ correspondencia\ del\ tipo\ i\ correctas}{total\ de\ unidades\ de\ correspondencia\ del\ tipo\ i} \tag{5.2}$$

Donde i se refiere a alguno de los casos posibles de alineación entre oraciones, por ejemplo: 1:1, 1:2, 2:1,...

De manera análoga con la ecuación 5.1 se considera a una ***unidad de correspondencia de cierto tipo correcta*** como aquella que coincide elemento a elemento con alguna unidad del mismo tipo contenida en el *gold standar*.

5.3. Evaluación del método

La evaluación del sistema se llevará a cabo comparando cada una de las unidades de correspondencia expresadas en la alineación que devuelve el método con cada una de las unidades de correspondencia contenidas en el gold standar.

En las tablas 5.5 y 5.4 se muestran la efectividad obtenida por el método propuesto para cada sección del corpus de prueba y la efectividad por en cada tipo de alineación presente en el corpus.

Tabla 5.4: Efectividad del método en el corpus *Otra vuelta a la tuerca*

Sección	Efectividad	Tipo de alineación	Aciertos	Efectividad por tipo
capítulo 1–2	94 %	1:1	290	96.66 %
		1:2	10	83.33 %
		2:1	6	75 %
		2:2	1	100 %
		1:∅	1	100 %
		∅:1	1	100 %

Tabla 5.5: Efectividad del método en el corpus *Las aventuras de Sherlock Holmes*

Sección	Efectividad	Tipo de alineación	Aciertos	Efectividad por tipo
capítulo 1–4	95.624 %	1:1	1995	96.66 %
		1:2	54	83.07 %
		2:1	57	89.06 %
		2:2	2	50 %
		3:1	1	100 %
		1:3	2	50 %
		1:∅	6	54.54 %
		∅:1	4	57.14 %

Los errores que se presentaron en los experimentos, se deben a:

- Un mal procesamiento de la oración: hace referencia a palabras sobre las que no se obtiene información o se obtiene información errónea. Se presenta en los siguientes casos:
  - La palabra se encuentra mal codificada.
  - El analizador morfológico no reconoce la palabra.
  - Se clasifica mal a la palabra.
  - No existe el lema de la palabra como entrada en el diccionario bilingüe.
  - La información que regresa el diccionario bilingüe sobre una palabra es errónea.
- Mala elección de una unidad de correspondencia: hace referencia a la propagación del error ante una mala elección en las etapas siguientes. Dado que el espacio de búsqueda ha sido segmentado la propagación del error se encuentra acotada por los limites del segmento, pero existen segmentos que de un tamaño considerable especialmente en partes del texto donde se encuentran diálogos.
- Naturaleza de la traducción: se refiere a aquellos casos en los que la traducción de una palabra se realiza a través de un conjunto de palabras, al uso de frases propias de cada lenguaje o a adecuaciones por parte del traductor. En la tabla 5.6 se muestran algunos ejemplos.

Tabla 5.6: *Ejemplos de traducciones no técnicas*

Texto en inglés	**Texto en español**	**Fenómeno**
Do not *join in it*	No *intervenga*	correspondencia múltiple
Oh, dear!	*¡Valgame Dios!*	frases propias
But the *note* itself	Pero en cuanto a la *carta* en si	adecuaciones

5.3.1. Comparación con otros métodos

Los resultados obtenidos con el método propuesto se compararon contra los resultados obtenidos por un alineador a nivel de oraciones estadístico llamado *Vanilla aligner* desarrollado por Danielsson & Riddings [1994].

Vanilla aligner se basa en el trabajo de Gale & Church [1991] y contempla los siguientes tipos de alineación entre oraciones: $1:1$, $1:\varnothing$, $\varnothing:1$, $1:2$, $2:1$ y $2:2$. El sistema requiere como entrada una versión segmentada a nivel de oraciones del corpus a alinear con una codificación especial, entonces como entrada se utiliza la misma segmentación del método propuesto (véase sección 4.3.1).

Se realizó la misma prueba que con el método propuesto y se evaluó solamente la efectividad de las alineaciones encontradas para cada uno del corpus de prueba. En las tabla 5.7 se muestra un comparativo de la eficiencia obtenida por el método propuesto y la obtenida por el sistema *Vanilla aligner*.

Tabla 5.7: Comparativo de eficiencia obtenida por el método propuesto y Vanilla aligner

Título	Método propuesto	*Vanilla aligner*
Otra vuelta a la tuerca	96.78 %	93.01 %
Las aventuras de Sherlock Holmes	95.624 %	90.66 %

Capítulo 6

Conclusiones

6.1. Conclusiones y trabajo futuro

En este trabajo de tesis, se desarrolló un método para la alineación de textos paralelos a nivel de oraciones escritos en los idiomas español e inglés, el cuál utiliza información léxica y estadística bajo un enfoque de programación dinámica.

El método utiliza la información léxica contenida en un diccionario bilingüe español–inglés de propósito general restringido (incompleto), así como, el número de elementos significativos y la longitud de la oración medida en términos de caracteres.

El método se implementó bajo el entorno de desarrollo C++ Builder y se utilizó para alinear un corpus de prueba conformado por algunos capítulos de novelas de novelas de ciencia ficción. El tipo de textos utilizados fueron elegidos por tener una mayor frecuencia de omisiones, inserciones y alineaciones múltiples, casos que se consideran difíciles dentro del ámbito de alineación de corpus paralelo a nivel de oraciones.

Los resultados obtenidos fueron comparados con alineaciones realizadas por un humano y por un sistema que utiliza una metodología estadística (Vanilla aligner). Los resultados mostraron que el método obtuvo una efectividad superior al 90% en un corpus literario, mostrando un buen desempeño en casos de alineaciones multiples, omisiones e inserciones, así como, un desempeño mayor respecto a una metodología estadística.

La experiencia presentada deja entre ver que el uso de información léxica, particularmente la proporcionada por un diccionario bilingüe, es un recurso que permite obtener un buen desempeño en textos que presentan traducciones literarias, al proveer cierto nivel de certidumbre al momento de elegir una configuración de correspondencia para
determinado grupo de oraciones (unidades de correspondencia).

El uso de un diccionario bilingüe como recurso léxico puede proveer a distintos métodos enfocados a la alineación de corpus paralelo de cierto cierto nivel de robustez ante diferentes tipos de textos y al mismo tiempo ser un recurso accesible que no requiere ser exhaustivamente completo ni especializado, que puede ser adquirido de manera gratuita a través de la Web.

El sistema implementado fue utilizado para compilar un corpus paralelos español–inglés alineado a nivel de oraciones de dos novelas literarias: Las aventuras de Sherlock Holmes de Arthur Conan Doyle con una extensión de 12 capítulos (6662 oraciones del texto en inglés y 6748 oraciones del texto en español) y Otra vuelta de tuerca de Henry James con una extensión de 24 capítulos (2557 oraciones del texto en inglés y 2598 oraciones del texto es español).

Como trabajo futuro se plantean las siguientes acciones:

- Evaluar la efectividad del método en diferentes tipos de textos: institucionales, técnicos y científicos.
- Incluir un nuevo recurso léxico del que se puedan obtener hiperónimo y sinónimos de las palabras con el fin de obtener una mejor evaluación de la correspondencia entre oraciones.
- Plantear un algoritmo se segmentación de oraciones más exacto para el español e inglés.
- Incluir los casos en que la traducción de una palabra comprenda un conjunto de palabras o viceversa del español e inglés.

6.2. Aportaciones

6.2.1. Aportaciones científicas

Las aportaciones científicas de este trabajo son:

- Desarrollo de una método para la alineación de textos paralelos inglés–español que utiliza recursos lingüísticos (diccionario bilingüe inglés–español no especializado y analizador morfológico para español e inglés).
- Propuesta de una función de similitud para la alineación de oraciones de textos en español e inglés.
- Algoritmo de segmentación textos en oraciones para el español e inglés.
- Evaluación del método propuesto.

6.2.2. Aportaciones técnicas

Las aportaciones técnicas de este trabajo fueron:

- Implementación un método de alineación de corpus paralelo español–inglés a nivel de oraciones.
- Compilación un corpus de prueba paralelo español–inglés alineado a nivel de oraciones de textos literarios.
- Implementación de un algoritmo de segmentación de textos en oraciones para los idiomas español e inglés.

- Implementación de una representación estructurada con metadatos de un corpus segmentado a nivel de oraciones donde se incluyen las traducciones de cada palabra de acuerdo con un diccionario bilingüe.

Bibliografía

Bolshakov, I. A., Galicia-Haro, S. N. & Gelbukh, A. [2003], 'Stable coordinated pairs in text processing', *Václav Matousek, Pavel Mautner (Eds.) Text, Speech and Dialogue (TSD-2003: 6th International Conference), Ceske Budejovice, Czech Republic, September, 2003. Lecture Notes in Artificial Intelligence (indexed by SCIE), Springer-Verlag* pp. 27–34.

Brown, P. F., Cocke, J., Pietra, S. A. D., Pietra, V. J. D., Jelinek, F., Lafferty, J. D., Mercer, R. L. & Roossin, P. S. [1990], 'A statistical approach to machine translation', *Comput. Linguist.* **16**(2), 79–85.

Brown, P. F., DellaPietra, S. A., DellaPietra, V. J. & Mercer, R. L. [1991a], 'Word sense disambiguation using statistical methods', *In Poceedings 29ht Annual Meeting of the ACL* pp. 265–270. Berkley, Ca, June 1991.

Brown, P., Lai, J. & Mercer, R. [1991b], 'Aligning sentences in parallel corpora', *In Poceedings 29ht Annual Meeting of the ACL* pp. 169–176. Berkley, Ca, June 1991.

Callison-Burch, C., Koehn, P. & Osborne, M. [2006], 'Improved statistical machine translation using paraphrases', *In Proceedings NAACL-2006* .

Chan, Y. S., Wang, B. & Ng, H. T. [2003], 'Exploiting parallel texts for word sense disambiguation: An empirical study', *In Poceedings of ACL-03* pp. 455–462. Sapporo, Japan.

Chen, S. F. [1993], 'Aligning sentences in bilingual corpora using lexical information', *In Proceedings of ACL-93* pp. 9–16.

Collier, N., Ono, K. & Hirakawa, H. [1998], 'An experimental in hybrid dictionary and statistical sentence alignment' , *In Proceedings of the 17th international conference on Computational Linguistics* pp. 268–274.

Danielsson, P. & Riddings, D. [1994], Practical presentation of a vanilla aligner, *in* 'Presentation held at the TELRI Workshop in Alignment and EXplotation of TExts', pp. 1–2.

Gale, W. A. & Church, K. W. [1991], A program for aligning sentences in bilingual corpora, *in* 'Proceedings of the 29th annual meeting on Association for Computational Linguistics', Association for Computational Linguistics, Morristown, NJ, USA, pp. 177–184.

Gautam, M. & Sinha, R. [2007], 'A hybrid approach to sentence alignment using genetic algorithm', *International Conference on Computing: Theory and Applications* **0**, 480–484.

Gelbukh, A. & Sidorov, G. [2006*a*], 'Alignment of paragraphs in bilingual texts using bilingual dictionaries and dynamic programming', *Lecture Notes in Computer Science, Springer-Verlag* (4225), 824–833.

Gelbukh, A. & Sidorov, G. [2006*b*], 'Aproach to construction of automatic morphological analysis systems for inflective languages with little effort', *Lecture Notes in Computer Science, Springer-Verlag* (2588), 215–220.

Haruno, M. & Yamazaki, T. [1997], 'High-performance bilingual text alignment using statistical and dictionary information', *In Proceedings of the 17th international conference on Computational Linguistics* pp. 1–14.

Hays, D. [1963], 'Research procedures in machine translation.', *In Paul L. Garvin.(Ed.)Natural Language and the Computer(University of Californea Engineering and Sciences Extension Series). McGraw Hill*. New York.

Isahara, H. & Haruno, M. [2000], Japanese-english aligned bilingual corpora. Netherlands.

Jones, T. [2008], *Artificial Intelligence*, Infinity Science Press LLC. pp. 21-48.

Kay, M. & Röscheisen, M. [1993], 'Text-translation alignment', *Comput. Linguist.* **19**(1), 121–142.

Kim, S. D., Zhang, B.-T. & Kim, Y. T. [2000], Reducing parsing complexity by intra-sentence segmentation based on maximum entropy model, *in* 'Proceedings of the 2000 Joint SIGDAT conference on Empirical methods in natural language processing and very large corpora: held in conjunction with the 38th Annual Meeting of the Association for Computational Linguistics - Volume 13', EMNLP '00, Association for Computational Linguistics, Stroudsburg, PA, USA, pp. 164–171.

King, G. W. [1956], 'Stochastic methods of mechanical translation.', *In the journal Mechanical Translation. MIT PRESS* volume 3.

Kirby, S. & Hurford, J. [1997], 'Learning, culture and evolution inthe origin of linguistic constraints', *In Fourth European Conference on Artificial Life. MIT Press* .

Klavans, J. & Tzoukermann, E. [1990], 'The bicord system', *In COLING-90* pp. 174–179. Helsinki, Finland, Ausgust 1990.

Koutsoudas, A. & Humecky, A. [1957], *Ambiguity of syntactic function resolved by linear context*, Word. 13(3), pp. 403-414.

Li, W.-C., Pei, T., Lee, B.-H. & Chiou, C.-F. [1990], Parsing long english sentences with pattern rules, *in* 'Proceedings of the 13th conference on Computational linguistics - Volume 3', COLING '90, Association for Computational Linguistics, Stroudsburg, PA, USA, pp. 410–412.

Michalewicz, Z. & Fogel, D. [1998], *How to solve it: Modern Heuristics*, Springer. pp. 35-48.

Nemhauser, G. [1966], *Introduction to Dynamic Programming*, Wiley.

Nerbonne, J. [2000], 'Parallel texts in computer-assisted language learning.', *In Véronis,J.(Ed.),Parallel Text Processing. Dordrecht: Kluwer Academic Publishers* . Netherlands.

Oliphant, M. [1996], 'The dilemma of saussurean communication', *Biosystems* 37(1-2), 31–38.

Papageorgiou, H., Cranias, L. & Piperidis, S. [1994], Automatic alignment in parallel corpora, *in* 'Proceedings of the 32nd annual meeting on Association for Computational Linguistics', ACL '94, Association for Computational Linguistics, Stroudsburg, PA, USA, pp. 334–336.

Pierre, I. [1991], 'Une nouvelle génération d'aides à la traduction et la terminologie', *presented at the Terminologie et documentation colloquium* . Hull, October 1991.

Sato, S. [1992], 'Ctm: an example-based translation aid system', *In Proceedings 14th COLING* pp. 1259–1263.

Sidorov, G., Gelbukh, A. & Velázquez, F. [2002], Agme: Un sistema de análisis y generación de la morfología del español, *in* 'Proceedings of the Multilingual Information Access and Natural Language Processing, Intenational Workshop at IBERAMIA', Julio Gonzalo, Anselmo Peñas, Antonio Ferrández Eds., Sevilla, Spain, pp. 1–6.

Simard, M., Foster, G. & Isabelle, P. [1992], 'Using cognates to align sentences in bilingual corpora', *TMI–1992* pp. 67–81.

Singh, S. [1999], *The Code Book. The Science of Secrecy from Ancient Egypt to Quantum Cryptography*, Anchor Boks. United States of America, pp. 205-217.

Singh, S., McEnery, T. & Baker, P. [2000], Building a parallel corpus of english/panjabi. Netherlands.

Trujillo, A. [1999], *Tranlation Engines: Techniques for Machine Translation*, Springer. Inglaterra, pp. 3-7,57-82.

Véronis, J. [2000], 'A survey of parallel text processing: from the rosetta stone to the information society.', *In Véronis,J.(Ed.),Parallel Text Processing. Dordrecht: Kluwer Academic Publishers* . Netherlands.

Véronis, J. & Langlais, P. [2000], Evaluation of parallel text alignment systems. the arcade project. Netherlands.

Warwick, S. & Rusell, G. [1990], 'Bilingual concordancing and bilingual lexicography', *In EURALEX 4th International Congress* . Málaga, Spain.

Weaver, W. [1955], 'Translation.', *In William N. Locke and A. D. Booth(Eds.),Machine Translation of Languages: Fourteen Essays. jointly published by The Technology Press of Massachusetts Institute of Technology, John Wiley (New York) nad Chapman & Hall (London)* .

Wu, D. [1994], Aligning a parallel english-chinese corpus statistically with lexical criteria, *in* 'Proceedings of the 32nd annual meeting on Association for Computational Linguistics', Association for Computational Linguistics, Morristown, NJ, USA, pp. 80–87.

Xiong, H., Xu, W., Mi, H., Liu, Y. & Liu, Q. [2009], Sub-sentence division for tree-based machine translation, *in* 'Proceedings of the ACL-IJCNLP 2009 Conference Short Papers', ACLShort '09, Association for Computational Linguistics, Stroudsburg, PA, USA, pp. 137–140.

Anexo A

Ejemplo de alineación

A.1. Corpus de prueba

A.1.1. Texto en inglés

Sir Arthur Conan Doyle

The adventures of Sherlock Holmes

A Scandal in Bohemia

To Sherlock Holmes she is always the woman. I have seldom heard him mention her under any other name. In his eyes she eclipses and predominates the whole of her sex. It was not that he felt any emotion akin to love for Irene Adler. All emotions, and that one particularly, were abhorrent to his cold, precise but admirably balanced mind. He was, I take it, the most perfect reasoning and observing machine that the world has seen, but as a lover he would have placed himself in a false position. He never spoke of the softer passions, save with a gibe and a sneer. They were admirable things for the observer – excellent for drawing the veil from men´s motives and actions. But for the trained reasoner to admit such intrusions into his own delicate and finely adjusted temperament was to introduce a distracting factor which might throw a doubt upon all his mental results. Grit in a sensitive instrument, or a crack in one of his own high power lenses, would not be more disturbing than a strong emotion in a nature such as his. And yet there was but one woman to him, and that woman was the late Irene Adler, of dubious and questionable memory.

I had seen little of Holmes lately. My marriage had drifted us away from each other. My own complete happiness, and the home centred interests which rise up around the man who first finds himself master of his own establishment, were sufficient to absorb all my attention, while Holmes, who loathed every form of society with his whole Bohemian soul, remained in our lodgings in Baker Street, buried among his old books, and alternating from week to week between cocaine and ambition, the drowsiness of the drug, and the fierce energy of his own keen nature. He was still, as ever, deeply attracted by the study of crime, and occupied his immense faculties and extraordinary powers of observation in following out those clues, and clearing up those mysteries which had been abandoned as hopeless by the official police. From time to time I heard some vague account of his doings: of his summons to Odessa in the case of the Trepoff murder, of his clearing up of the singular tragedy of the Atkinson brothers at Trincomalee, and finally of the mission which he had accomplished so delicately and successfully for the reigning family of Holland. Beyond these signs of his activity, however, which I merely shared with all the readers of the daily press, I knew little of my former friend and companion.

One night – it was on the twentieth of March, 1888 – I was returning from a journey to a patient (for I had now returned to civil practice), when my way led me through Baker Street. As I passed the well remembered door, which must always be associated in my mind with my wooing, and with the dark incidents of the Study in Scarlet, I was seized with a keen desire to see Holmes again, and to know how he was employing his extraordinary powers. His rooms were brilliantly lit, and, even as I looked up, I saw his tall, spare figure pass twice in a dark silhouette against the blind. He was pacing the room swiftly, eagerly, with his head sunk upon his chest and his hands clasped behind him. To me, who knew his every mood and habit, his attitude and manner told their own story. He was at work again. He had risen out of his drug created dreams and was hot upon the scent of some new problem. I rang the bell and was shown up to the chamber which had formerly been in part my own.

His manner was not effusive. It seldom was; but he was glad, I think, to see me. With hardly a word spoken, but with a kindly eye, he waved me to an armchair, threw across his case of cigars, and indicated a spirit case and a gasogene in the corner. Then he stood before the fire and looked me over in his singular introspective fashion.

"Wedlock suits you," he remarked. "I think, Watson, that you have put on seven and a half pounds since I saw you."

"Seven" I answered.

"Indeed, I should have thought a little more. Just a trifle more, I fancy, Watson. And in practice again, I observe. You did not tell me that you intended to go into harness."

A.1.2. Texto en español

Arthur Conan Doyle

Las aventuras de Sherlock Holmes

I. Escándalo en Bohemia

Para Sherlock Holmes, ella es siempre la mujer. Rara vez le oí mencionarla de otro modo. A sus ojos, ella eclipsa y domina a todo su sexo. Y no es que sintiera por Irene Adler nada parecido al amor. Todas las emociones, y en especial ésa, resultaban abominables para su inteligencia fría y precisa pero admirablemente equilibrada. Siempre lo he tenido por la máquina de observar y razonar más perfecta que ha conocido el mundo; pero como amante no habría sabido qué hacer. Jamás hablaba de las pasiones más tiernas, si no era con desprecio y sarcasmo. Eran cosas admirables para el observador, excelentes para levantar el velo que cubre los motivos y los actos de la gente. Pero para un razonador experto, admitir tales intrusiones en su delicado y bien ajustado temperamento equivalía a introducir un factor de distracción capaz de sembrar de dudas todos los resultados de su mente. Para un carácter como el suyo, una emoción fuerte resultaba tan perturbadora como la presencia de arena en un instrumento de precisión o la rotura de una de sus potentes lupas. Y sin embargo, existió para él una mujer, y esta mujer fue la difunta Irene Adler, de dudoso y cuestionable recuerdo.

Últimamente, yo había visto poco a Holmes. Mi matrimonio nos había apartado al uno del otro. Mi completa felicidad y los intereses hogareños que se despiertan en el hombre que por primera vez pone casa propia bastaban para absorber toda mi atención; mientras tanto, Holmes, que odiaba cualquier forma de vida social con toda la fuerza de su alma bohemia, permaneció en nuestros aposentos de Baker Street, sepultado entre sus viejos libros y alternando una semana de cocaína con otra de ambición, entre la modorra de la droga y la fiera energía de su intensa personalidad. Como siempre, le seguía atrayendo el estudio del crimen, y dedicaba sus inmensas facultades y extraordinarios poderes de observación a seguir pistas y aclarar misterios que la policía había abandonado por imposibles. De vez en cuando, me llegaba alguna vaga noticia de sus andanzas: su viaje a Odesa para intervenir en el caso del asesinato de Trepoff, el esclarecimiento de la extraña tragedia de los hermanos Atkinson en Trincomalee y, por último, la misión que tan discreta y eficazmente había llevado a cabo para la familia

real de Holanda. Sin embargo, aparte de estas señales de actividad, que yo me limitaba a compartir con todos los lectores de la prensa diaria, apenas sabía nada de mi antiguo amigo y compañero.

Una noche – –la del 20 de marzo de 1888– – volvía yo de visitar a un paciente (pues de nuevo estaba ejerciendo la medicina), cuando el camino me llevó por Baker Street. Al pasar frente a la puerta que tan bien recordaba, y que siempre estará asociada en mi mente con mi noviazgo y con los siniestros incidentes del Estudio en escarlata, se apoderó de mí un fuerte deseo de volver a ver a Holmes y saber en qué empleaba sus extraordinarios poderes. Sus habitaciones estaban completamente iluminadas, y al mirar hacia arriba vi pasar dos veces su figura alta y delgada, una oscura silueta en los visillos. Daba rápidas zancadas por la habitación, con aire ansioso, la cabeza hundida sobre el pecho y las manos juntas en la espalda. A mí, que conocía perfectamente sus hábitos y sus humores, su actitud y comportamiento me contaron toda una historia. Estaba trabajando otra vez. Había salido de los sueños inducidos por la droga y seguía de cerca el rastro de algún nuevo problema. Tiré de la campanilla y me condujeron a la habitación que, en parte, había sido mía. No estuvo muy efusivo; rara vez lo estaba, pero creo que se alegró de verme. Sin apenas pronunciar palabra, pero con una mirada cariñosa, me indicó una butaca, me arrojó su caja de cigarros, y señaló una botella de licor y un sifón que había en la esquina. Luego se plantó delante del fuego y me miró de aquella manera suya tan ensimismada.

– –El matrimonio le sienta bien – –comentó– –. Yo diría, Watson, que ha engordado usted siete libras y media desde la última vez que le vi.

– –Siete – –respondí.

– –La verdad, yo diría que algo más. Sólo un poquito más, me parece a mí, Watson. Y veo que está ejerciendo de nuevo. No me dijo que se proponía volver a su profesión.

A.2. Segmentación del corpus en oraciones

A.2.1. Segmentación del texto en inglés

Tabla A.1: Ejemplo de segmentación de un texto en inglés en oraciones

No. oración	Texto en inglés
1	Sir Arthur Conan Doyle
2	The adventures of Sherlock Holmes
3	A Scandal in Bohemia
4	To Sherlock Holmes she is always the woman.
5	I have seldom heard him mention her under any other name.
6	In his eyes she eclipses and predominates the whole of her sex.
7	It was not that he felt any emotion akin to love for Irene Adler.
8	All emotions, and that one particularly, were abhorrent to his cold, precise but admirably balanced mind.
9	He was, I take it, the most perfect reasoning and observing machine that the world has seen, but as a lover he would have placed himself in a false position.
10	He never spoke of the softer passions, save with a gibe and a sneer.
11	They were admirable things for the observer – excellent for drawing the veil from men´s motives and actions.
12	But for the trained reasoner to admit such intrusions into his own delicate and finely adjusted temperament was to introduce a distracting factor which might throw a doubt upon all his mental results.
13	Grit in a sensitive instrument, or a crack in one of his own high power lenses, would not be more disturbing than a strong emotion in a nature such as his.
14	And yet there was but one woman to him, and that woman was the late Irene Adler, of dubious and questionable memory.
15	I had seen little of Holmes lately.
16	My marriage had drifted us away from each other.

Tabla A.1: Ejemplo de segmentación de un texto en inglés en oraciones(continuación)

No. oración	Texto en inglés
17	My own complete happiness, and the home centred interests which rise up around the man who first finds himself master of his own establishment, were sufficient to absorb all my attention, while Holmes, who loathed every form of society with his whole Bohemian soul, remained in our lodgings in Baker Street, buried among his old books, and alternating from week to week between cocaine and ambition, the drowsiness of the drug, and the fierce energy of his own keen nature.
18	He was still, as ever, deeply attracted by the study of crime, and occupied his immense faculties and extraordinary powers of observation in following out those clues, and clearing up those mysteries which had been abandoned as hopeless by the official police.
19	From time to time I heard some vague account of his doings: of his summons to Odessa in the case of the Trepoff murder, of his clearing up of the singular tragedy of the Atkinson brothers at Trincomalee, and finally of the mission which he had accomplished so delicately and successfully for the reigning family of Holland.
20	Beyond these signs of his activity, however, which I merely shared with all the readers of the daily press, I knew little of my former friend and companion.
21	One night – it was on the twentieth of March, 1888 – I was returning from a journey to a patient (for I had now returned to civil practice), when my way led me through Baker Street.
22	As I passed the well remembered door, which must always be associated in my mind with my wooing, and with the dark incidents of the Study in Scarlet, I was seized with a keen desire to see Holmes again, and to know how he was employing his extraordinary powers.

Tabla A.1: Ejemplo de segmentación de un texto en inglés en oraciones(continuación)

No. oración	Texto en inglés
23	His rooms were brilliantly lit, and, even as I looked up, I saw his tall, spare figure pass twice in a dark silhouette against the blind.
24	He was pacing the room swiftly, eagerly, with his head sunk upon his chest and his hands clasped behind him.
25	To me, who knew his every mood and habit, his attitude and manner told their own story.
26	He was at work again.
27	He had risen out of his drug created dreams and was hot upon the scent of some new problem.
28	I rang the bell and was shown up to the chamber which had formerly been in part my own.
29	His manner was not effusive.
30	It seldom was; but he was glad, I think, to see me.
31	With hardly a word spoken, but with a kindly eye, he waved me to an armchair, threw across his case of cigars, and indicated a spirit case and a gasogene in the corner.
32	Then he stood before the fire and looked me over in his singular introspective fashion.
33	Wedlock suits you, he remarked.
34	I think, Watson, that you have put on seven and a half pounds since I saw you.
35	Seven I answered.
36	Indeed, I should have thought a little more.
37	Just a trifle more, I fancy, Watson.
38	And in practice again, I observe.
39	You did not tell me that you intended to go into harness.

A.2.2. Segmentación del texto en español

Tabla A.2: Ejemplo de segmentación de un texto en español en oraciones

No. oración	Texto en español
1	Arthur Conan Doyle
2	Las aventuras de Sherlock Holmes
3	I.
4	Escándalo en Bohemia
5	Para Sherlock Holmes, ella es siempre la mujer.
6	Rara vez le oí mencionarla de otro modo.
7	A sus ojos, ella eclipsa y domina a todo su sexo.
8	Y no es que sintiera por Irene Adler nada parecido al amor.
9	Todas las emociones, y en especial ésa, resultaban abominables para su inteligencia fría y precisa pero admirablemente equilibrada.
10	Siempre lo he tenido por la máquina de observar y razonar más perfecta que ha conocido el mundo; pero como amante no habría sabido qué hacer.
11	Jamás hablaba de las pasiones más tiernas, si no era con desprecio y sarcasmo.
12	Eran cosas admirables para el observador, excelentes para levantar el velo que cubre los motivos y los actos de la gente.
13	Pero para un razonador experto, admitir tales intrusiones en su delicado y bien ajustado temperamento equivalía a introducir un factor de distracción capaz de sembrar de dudas todos los resultados de su mente.
14	Para un carácter como el suyo, una emoción fuerte resultaba tan perturbadora como la presencia de arena en un instrumento de precisión o la rotura de una de sus potentes lupas.
15	Y sin embargo, existió para él una mujer, y esta mujer fue la difunta Irene Adler, de dudoso y cuestionable recuerdo.
16	Últimamente, yo había visto poco a Holmes.
17	Mi matrimonio nos había apartado al uno del otro.

Tabla A.2: Ejemplo de segmentación de un texto en español en oraciones(continuación)

No. oración	Texto en español
18	Mi completa felicidad y los intereses hogareños que se despiertan en el hombre que por primera vez pone casa propia bastaban para absorber toda mi atención; mientras tanto, Holmes, que odiaba cualquier forma de vida social con toda la fuerza de su alma bohemia, permaneció en nuestros aposentos de Baker Street, sepultado entre sus viejos libros y alternando una semana de cocaína con otra de ambición, entre la modorra de la droga y la fiera energía de su intensa personalidad.
19	Como siempre, le seguía atrayendo el estudio del crimen, y dedicaba sus inmensas facultades y extraordinarios poderes de observación a seguir pistas y aclarar misterios que la policía había abandonado por imposibles.
20	De vez en cuando, me llegaba alguna vaga noticia de sus andanzas: su viaje a Odesa para intervenir en el caso del asesinato de Trepoff, el esclarecimiento de la extraña tragedia de los hermanos Atkinson en Trincomalee y, por último, la misión que tan discreta y eficazmente había llevado a cabo para la familia real de Holanda.
21	Sin embargo, aparte de estas señales de actividad, que yo me limitaba a compartir con todos los lectores de la prensa diaria, apenas sabía nada de mi antiguo amigo y compañero.
22	Una noche – –la del 20 de marzo de 1888– – volvía yo de visitar a un paciente (pues de nuevo estaba ejerciendo la medicina), cuando el camino me llevó por Baker Street.
23	Al pasar frente a la puerta que tan bien recordaba, y que siempre estará asociada en mi mente con mi noviazgo y con los siniestros incidentes del Estudio en escarlata, se apoderó de mí un fuerte deseo de volver a ver a Holmes y saber en qué empleaba sus extraordinarios poderes.
24	Sus habitaciones estaban completamente iluminadas, y al mirar hacia arriba vi pasar dos veces su figura alta y delgada, una oscura silueta en los visillos.

Tabla A.2: Ejemplo de segmentación de un texto en español en oraciones(continuación)

No. oración	Texto en español
25	Daba rápidas zancadas por la habitación, con aire ansioso, la cabeza hundida sobre el pecho y las manos juntas en la espalda.
26	A mí, que conocía perfectamente sus hábitos y sus humores, su actitud y comportamiento me contaron toda una historia.
27	Estaba trabajando otra vez.
28	Había salido de los sueños inducidos por la droga y seguía de cerca el rastro de algún nuevo problema.
29	Tiré de la campanilla y me condujeron a la habitación que, en parte, había sido mía.
30	No estuvo muy efusivo; rara vez lo estaba, pero creo que se alegró de verme.
31	Sin apenas pronunciar palabra, pero con una mirada cariñosa, me indicó una butaca, me arrojó su caja de cigarros, y señaló una botella de licor y un sifón que había en la esquina.
32	Luego se plantó delante del fuego y me miró de aquella manera suya tan ensimismada.
33	– –El matrimonio le sienta bien – –comentó– –.
34	Yo diría, Watson, que ha engordado usted siete libras y media desde la última vez que le vi.
35	– –Siete – –respondí.
36	– –La verdad, yo diría que algo más.
37	Sólo un poquito más, me parece a mí, Watson.
38	Y veo que está ejerciendo de nuevo.
39	No me dijo que se proponía volver a su profesión.

A.3. Procesamiento del corpus

A.3.1. Texto en inglés

En la tabla A.3 se muestra solamente el procesamiento realizado a las primeras 4 oraciones del texto de ejemplo.

Tabla A.3: Ejemplo de procesamiento de un texto en inglés

<Preprocessing>
<text file="holmes25_eng.txt" num_sentence="27" num_char="2703" language="English">
<sentence id_sentence="1" num_words="4" mean_words="4" char="19" language="English">
<token char="Sir" class="normal">
<type char="sir">
<translation>señor</translation>
<translation>sir</translation>
</type>
</token>
<token char="Arthur" class="nombre_propio">
<type char="arthur">
<translation>arthur</translation>
</type>
</token>
<token char="Conan" class="nombre_propio">
<type char="conan">
<translation>conan</translation>
</type>
</token>
<token char="Doyle" class="nombre_propio">
<type char="doyle">
<translation>doyle</translation>
</type>
</token>
</sentence>
<sentence id_sentence="2" num_words="5" mean_words="3" char="29" language="English">
<token char="The" class="palabra_auxiliar">
<type char="the"/>
</token>
<token char="adventures" class="normal">
<type char="adventure">
<translation>aventura</translation>
</type>
</token>
<token char="of" class="palabra_auxiliar">
<type char="of"/>

Tabla A.3: Ejemplo de procesamiento de un texto en inglés (continuación)

```
</token>
<token char="Sherlock" class="nombre_propio">
<type char="sherlock">
<translation>sherlock</translation>
</type>
</token>
<token char="Holmes" class="nombre_propio">
<type char="holmes">
<translation>holmes</translation>
</type>
</token>
</sentence>
<sentence id_sentence="3" num_words="4" mean_words="2" char="17" language="English">
<token char="A" class="palabra_auxiliar">
<type char="a"/>
</token>
<token char="Scandal" class="normal">
<type char="scandal">
<translation>escándalo</translation>
<translation>vergüenza</translation>
<translation>chismes</translation>
<translation>chismorreo</translation>
</type>
</token>
<token char="in" class="palabra_auxiliar">
<type char="in"/>
</token>
<token char="Bohemia" class="nombre_propio">
<type char="bohemia">
<translation>bohemia</translation>
</type>
</token>
</sentence>
<sentence id_sentence="4" num_words="8" mean_words="4" char="35" language="English">
<token char="To" class="palabra_auxiliar">
<type char="to"/>
</token>
```

Tabla A.3: Ejemplo de procesamiento de un texto en inglés (continuación)

```
<token char="Sherlock" class="nombre_propio">
<type char="sherlock">
<translation>sherlock</translation>
</type>
</token>
<token char="Holmes" class="nombre_propio">
<type char="holmes">
<translation>holmes</translation>
</type>
</token>
<token char="she" class="palabra_auxiliar">
<type char="she"/>
</token>
<token char="is" class="palabra_auxiliar">
<type char="is"/>
</token>
<token char="always" class="normal">
<type char="always">
<translation>siempre</translation>
</type>
</token>
<token char="the" class="palabra_auxiliar">
<type char="the"/>
</token>
<token char="woman" class="normal">
<type char="woman">
<translation>mujer</translation>
<translation>señora</translation>
</type>
</token>
</sentence>
```

A.3.2. Texto en español

En la tabla A.4 se muestra solamente el procesamiento realizado a las primeras 5 oraciones del texto de ejemplo.

Tabla A.4: Ejemplo de procesamiento de un texto en español

```
<Preproccessing>
<text file="holmes25_spa.txt" num_sentence="28" num_char="2828" language="Spanish">
<sentence id_sentence="1" num_words="3" mean_words="3" char="16" language="Spanish">
<token char="Arthur" class="nombre_propio">
<type char="arthur">
<translation>arthur</translation>
</type>
</token>
<token char="Conan" class="nombre_propio">
<type char="conan">
<translation>conan</translation>
</type>
</token>
<token char="Doyle" class="nombre_propio">
<type char="doyle">
<translation>doyle</translation>
</type>
</token>
</sentence>
<sentence id_sentence="2" num_words="5" mean_words="3" char="28" language="Spanish">
<token char="Las" class="palabra_auxiliar">
<type char="las"/>
</token>
<token char="aventuras" class="normal">
<type char="aventura">
<translation>adventure</translation>
<translation>hazard</translation>
<translation>risk</translation>
<translation>affair</translation>
</type>
<type char="aventurar">
<translation>hazard</translation>
<translation>risk</translation>
<translation>venture</translation>
<translation>dare</translation>
</type>
```

Tabla A.4: Ejemplo de procesamiento de un texto en español (continuación)

```
</token>
<token char="de" class="palabra_auxiliar">
<type char="de"/>
</token>
<token char="Sherlock" class="nombre_propio">
<type char="sherlock">
<translation>sherlock</translation>
</type>
</token>
<token char="Holmes" class="nombre_propio">
<type char="holmes">
<translation>holmes</translation>
</type>
</token>
</sentence>
<sentence id_sentence="3" num_words="1" mean_words="1" char="1" language="Spanish">
<token char="1" class="numero">
<type char="1">
<translation>1</translation>
</type>
</token>
</sentence>
<sentence id_sentence="4" num_words="3" mean_words="2" char="18" language="Spanish">
<token char="Escándalo" class="normal">
<type char="escándalo">
<translation>scandal</translation>
<translation>racket</translation>
<translation>fuss</translation>
<translation>din</translation>
<translation>uproar</translation>
<translation>astonishment</translation>
<translation>shock</translation>
</type>
</token>
<token char="en" class="palabra_auxiliar">
<type char="en"/>
</token>
```

Tabla A.4: Ejemplo de procesamiento de un texto en español (continuación)


```
<token char="Bohemia" class="normal">
<type char="bohemia">
<translation>Bohemia</translation>
</type>
<type char="bohemio">
<translation>bohemian</translation>
<translation>Bohemian</translation>
</type>
</token>
</sentence>
<sentence id_sentence="5" num_words="8" mean_words="3" char="38" language="Spanish">
<token char="Para" class="palabra_auxiliar">
<type char="para"/>
</token>
<token char="Sherlock" class="nombre_propio">
<type char="sherlock">
<translation>sherlock</translation>
</type>
</token>
<token char="Holmes" class="nombre_propio">
<type char="holmes">
<translation>holmes</translation>
</type>
</token>
<token char="ella" class="palabra_auxiliar">
<type char="ella"/>
</token>
<token char="es" class="palabra_auxiliar">
<type char="es"/>
</token>
<token char="siempre" class="palabra_auxiliar">
<type char="siempre"/>
</token>
<token char="la" class="palabra_auxiliar">
<type char="la"/>
</token>
<token char="mujer" class="normal">
```

Tabla A.4: Ejemplo de procesamiento de un texto en español (continuación)

```
<type char="mujer">
<translation>woman</translation>
<translation>wife</translation>
</type>
</token>
</sentence>
```

A.4. Alineación

La alineación realizada por la implementación del método, se muestra en la tabla A.5 en ella se muestran los indices de las oraciones de que corresponden entre sí de cuerdo con las tablas A.1 y A.2.

Tabla A.5: Alineación de un texto de ejemplo

No. oración en inglés	No. de oración en español
1	1
2	2
vacio	3
3	4
4	5
5	6
6	7
7	8
8	9
9	10
10	11
11	12
12	13
13	14
14	15
15	16
16	17
17	18

Tabla A.5: Alineación de un texto de ejemplo (continuación)

No. oración en inglés	No. de oración en español
18	19
19	20
20	21
21	22
22	23
23	24
24	25
25	26
26	27
27	28
28	29
29,30	30
31	31
32	32
33	33
34	34
35	35
36	36
37	37
38	38
39	39