UNIVERSIDAD COMPLUTENSE DE MADRID

FACULTAD DE FILOLOGíA

GRADO EN LINGÜÍSTICA Y LENGUAS APLICADAS

TRABAJO DE FIN DE GRADO

Segmentación automática de palabras en chino

Autor: Lin Ma

Tutora: Ana Fernández-Pampillón Cesteros

CURSO ACADÉMICO: 2017-2018

CONVOCATORIA: Septiembre

Índice

Resumen

La segmentación es una tarea importante en el procesamiento del lenguaje natural por ser el primer paso para poder analizar correctamente textos. La calidad del resultado de la segmentación va a afectar directamente a la eficiencia del procesamiento textual. La segmentación no es un proceso trivial, en general, y, en particular, presenta más dificultades en las lenguas que no segmentadas como el chino, la lengua con mayor número de hablantes en el mundo. Este trabajo, en primer lugar, ofrece una visión general de los conceptos básicos de la segmentación de palabras; en segundo lugar, revisa la segmentación automática en chino; en tercer lugar, se analizan las dificultades de la segmentación automática, en palabras, de textos chinos; y, finalmente, se presenta una propuesta de trabajo futuro y las conclusiones. Su objetivo es que pueda servir para ayudar a conocer y entender el chino desde la perspectiva de la segmentación y así motivar la realización de trabajos de mejora de la segmentación automática en chino.

Palabras claves:

Segmentación, chino, procesamiento de textos, Lingüística Computacional

1. Introducción

Con el desarrollo de internet, estamos viviendo en una era llena de información. La dependencia del ser humano con las computadoras se refleja en todos los aspectos de la vida. Las computadoras procesan enormes cantidades de información textual, por lo tanto, el uso del lenguaje natural para comunicarse con las computadoras, denominado Procesamiento del Lenguaje Natural (en adelante PLN) o Lingüística Computacional, se ha convertido en un tema de interés académico, económico y social. Yoav Goldberg escribió en su libro "Neural Network Methods for Natural Language Processing" que el PLN es un término colectivo que se refiere al procesamiento automático de los lenguajes humanos, lo que incluye tanto a los algoritmos que toman texto o audio producido por el humano como entrada, como los algoritmos que generan o producen texto o audio como salida (Yoav, 2017). En este trabajo nos vamos a ocupar solamente del procesamiento de textos, no de audio. Por otra parte, en el PLN, el proceso de procesar la entrada se llama comprensión o análisis del lenguaje natural (*Natural Language Understanding*), y el proceso de procesar la salida se llama generación de lenguaje natural (*Natural Language Generation*). La segmentación sólo se lleva a cabo en la comprensión del lenguaje natural. La comprensión de textos incluye diversas tareas entre las que destaca, por ser la base de las tareas posteriores, la tarea de segmentación. Para que la computadora comprenda perfectamente un texto, la segmentación automática de entrada tiene que ser correcta. La calidad del resultado de la segmentación va a afectar directamente a la eficiencia del procesamiento de entrada.

Actualmente existe un problema serio para las aplicaciones PLN, como los buscadores de internet o traductores automáticos, en relación al idioma chino por su dificultad para segmentarlo. Resolver este problema facilitaría la comprensión entre las cultura occidental y oriental. Por ello, en este trabajo nos proponemos buscar la respuesta a tres de los principales problemas con los cuales nos enfrentamos en la segmentación del idioma chino.

- ¿Por qué es difícil segmentar en chino?
- ¿Cómo segmentar las palabras en chino?
- ¿Cuáles son las dificultades en la segmentación del chino?

2. Objetivos del trabajo

Por todos es sabido, que el idioma chino es muy diferente a los idiomas occidentales. Entre las muchas diferencias, una de las más significativas, desde el punto de vista de la segmentación, es que no existe un espacio entre los caracteres. Esta diferencia hace que la segmentación automática de las palabras en chino no sea una tarea fácil para los ordenadores.

El objetivo de este trabajo es contribuir a conocer los algoritmos de segmentación automática de palabras en el idioma chino y determinar las dificultades de segmentación automática con el fin de ayudar a futuras investigaciones.

3. Metodología del trabajo

Este trabajo ha sido desarrollado en tres fases:

(1) En una primera fase se ha realizado una búsqueda de bibliografía en dos fuentes:

1. Bibliografía obtenida de la biblioteca de la Universidad Complutense de Madrid, en concreto de la facultad de Filología y a través de la base de datos bibliográfica de libros, artículos de la Universidad Complutense de Madrid.
2. Bibliografía obtenida de las paginas web de la Peking University Library (https://www.lib.pku.edu.cn/portal/) y de la pagina web de la Tsinghua University Library (http://www.lib.tsinghua.edu.cn/index.html), suscripción y descarga de artículos científicos de la revista Computer Knowledge and technology.

(2) En la segunda fase, se analizaron, tradujeron y sintetizaron las fuentes bibliográficas obtenidas de autores europeos y de autores chinos, centrando la información en la segmentación de palabras, modelos, algoritmos y problemática de los mismos.

(3) Finalmente, en la tercera fase, se ha realizado la redacción de este trabajo con el objetivo de identificar los problemas derivados de la segmentación de palabras en el idioma chino, la búsqueda de la solución y las vías futuras que deben seguir las investigaciones para solucionar este problema.

4. Segmentación automática

Desde el punto de vista lingüístico un texto se puede segmentar en las unidades lingüísticas como palabras, grupos sintácticos, frases, párrafos, etc. Desde el punto de vista digital, el texto electrónico es una secuencia de caracteres, algunos de los cuales son caracteres alfanuméricos, como las letras, los números, las puntuaciones y otros son tipografía, como espacios en blanco, nueva línea.

Antes de que se realice cualquier procesamiento de texto electrónico, el texto debe estar segmentado, al menos, en unidades lingüísticas con sentido, como palabras, frase, puntuación, números, alfanuméricos, etc. Este proceso de división en unidades lingüísticas se llama **tokenización** o **segmentación** (MITKOV, 2003). La segmentación es un análisis en el que la salida es el conjunto de las palabras del texto, las frases del texto u otro tipo de segmentos.

4.1 Segmentación automática de palabras

Segmentación de palabras significa segmentar el texto con una salida de un conjunto de las "palabras" del texto. Utilizamos las comillas de "palabras" aquí, porque no son siempre las palabras del diccionario, sino son las unidades separadas, llamadas tokens, con una significación propia. Por esta razón, el proceso para segmentar un texto en tokens de palabra se llama **tokenización** (MITKOV, 2003).

A modo de ilustración, en muchos idiomas europeos, como el español, hay un límite claro entre las palabras y el espacio. Estos idiomas solo necesitan usar el espacio como el límite de tokens, y luego eliminan los signos de puntuación al principio y al final, la segmentación ya ha logrado un éxito inicial. Esto no quiere decir que cada palabra separada por el espacio es un token de palabra. Por ejemplo, la frase "Hoy es uno de junio.", los tokens de palabra son "Hoy", "es", "uno de junio", "uno de junio" es un token con tres palabras, y cada una está separada por un espacio, pero esta fecha completa es un token de palabra, que tiene que estar junto. Además de la fecha que mencionamos anteriormente, las abreviaturas, los guiones, los números, los nombres propios, etc. todo estos dificultan la segmentación automática de palabras.

Hasta aquí, se puede preguntar cómo sabe un ordenador cuándo separar las palabras y cuándo

juntar las palabras. Tradicionalmente, las reglas de tokenización se escriben usando expresiones regulares que describen cómo hacer coincidir diferentes tipos de tokens como palabras, números, signos de puntuación, etc. **Expresión regular** es una expresión que describe un conjunto de cadenas (MITKOV, 2003). Por ejemplo, aquí tiene un ejemplo de expresión regular que coincide dos palabra que empiezan por letra mayúscula, y ellas está separando por espacio: " [A-Z]\w* [A-Z]\w* ". Si usas esta expresión en un programa con la frase "Me llamo Lin Ma.", encontrará un token de palabra que sea "Lin Ma". Entonces, si quiere una segmentación precisa de los tokens de palabra, solo tiene que poner la expresión regular correcta y ya está.

Por supuesto, no todos los idiomas tienen un límite claro entre palabras. Por ejemplo, los idiomas asiáticos como el chino, el coreano etc. no tienen espacio entre palabras. Por lo tanto, la segmentación automática de estos idiomas, que no tienen límite entre palabras es mucho más compleja. En la parte posterior de este trabajo, explicaré el algoritmo de segmentación automática de las palabras en chino..

4.2 Segmentación automática de frases

Segmentación de frases significa segmentar el texto con una salida de un conjunto de las frases del texto. La segmentación de frases es importante para muchas aplicaciones de procesamiento de lenguaje natural, por ejemplo; análisis sintáctico, traducción automática, resumen de documentos, etc. A diferencia del espacio como límite de palabras, el límite entre las frases son signos de puntuación, como el punto, el signo de exclamación o el signo de interrogación. Sin embargo, el punto también se puede usar como un decimal en números, o como una parte de abreviatura de inglés, por lo que también existen muchas dificultades en la segmentación de frases.

En la actualidad, el algoritmo más usado de la segmentación de frases es "período-espacio-letra con mayúscula [1] ". Sin embargo, la producción de este algoritmo no es muy buena, porque produce una tasa de error de aproximadamente 6,5 por ciento medida en el Brown corpus [2](MITKOV, 2003). Para perfeccionar el algoritmo "período-espacio-letra con mayúscula", puede listar previamente algunas abreviaturas comunes en inglés, como "Dr." o "Mr." en la

[1] "período-espacio-letra con mayúscula" marca todos los signos de punto, interrogación, y exclamación como al final de una frase, si ellos seguido de al menos un espacio en blanco y una letra mayúscula

[2] Compiled in the 1960s by Henry Kučera and W. Nelson Francisat Brown University, Providence, Rhode Island as a general corpus (text collection) in the field of corpus linguistics. It contains 500 samples of English-language text, totaling roughly one million words, compiled from works published in the United States in 1961.

programación, una vez que se encuentra estas abreviatura, aunque se cumplen las condiciones del algoritmo "período-espacio-letra con mayúscula", no se puede segmentar como una frase. En los siguientes puntos voy a enfocarme más en la segmentación de las palabras.

5. Conceptos de la segmentación automática de palabras en chino

La segmentación automática de palabras en chino es un caso especial dentro de la segmentación automática en general. Debido a la particularidad y la complejidad del idioma chino, la segmentación en palabras no es una tarea fácil. Antes de abordar el concepto de la segmentación de palabras en chino, vamos a introducir la diferencia entre lenguaje segmentado, como el español y no segmentado, como el chino.

5.1 Lenguaje segmentado y lenguaje no segmentado

Tanto en el idioma español como en el inglés que hemos mencionado anteriormente, los tokens de palabra ya están delimitados por espacios en blanco y signos de puntuación, el idioma en ellos se llama **lenguaje segmentado**. Entonces, al contrario del lenguaje segmentado, los tokens de palabras no tienen límites explícitos y se escriben directamente adyacentes uno u otro, el idioma de este tipo se llama **lenguaje no segmentado**(MITKOV, 2003). El idioma chino es un lenguaje no segmentado muy típico. Los caracteres, las frases y los párrafos de chino se pueden delimitar fácilmente mediante delimitadores explícitos, pero sólo las palabras no tienen ninguna forma de separador. En el idioma chino, una palabra es un token. ¿Pero cómo componer las palabras en chino? Principalmente, las palabras se componen de dobles caracteres o múltiples caracteres, aunque también puede ser de un carácter solo. Por ejemplo, la frase "我喜欢吃冰淇淋。", segmentaré en "我", "喜欢", "吃", "冰淇淋", con cuatro tokens. "我" es "yo", "喜欢" es el verbo "gustar", "吃" es el verbo "comer", y "冰淇淋" es "helado", el significado de la frase es "A mí me gusta comer helado.". "yo" y "comer" están compuesto de un carácter, "gustar" está compuesto de doble carácter, y el último, "helado" está compuesto de tres caracteres.

Con el idioma del chino, podemos ver fácilmente que la segmentación del lenguaje no segmentado es mucho más compleja que la del lenguaje segmentado. Su complejidad se refleja en que casi todos los lenguajes no segmentados son igual que el chino, algunos caracteres

individuales se pueden segmentar como un token, pero el resto deben combinar con otros caracteres y luego segmentar como un token. Pero cómo combinar, eso depende de la sintaxis de cada idioma.

5.2 Los caracteres de chino

Según la introducción del canon chino[3], los caracteres chinos son la unidad básica de la escritura China. Cada carácter es un individuo, tiene su propio significado, y se registra como un artículo en el diccionario. Pero no hay un número exacto de la cantidad de caracteres chino hasta el momento. Según las estadísticas, hay alrededor de 7000 caracteres. Entre ellos, hay alrededor de 1000 que son caracteres de uso común (La página web del canon chino).

Los caracteres chinos componen las palabras, las palabras componen las frases, y las frases componen los párrafo.

5.3 Las palabras de chino

Las unidades lingüísticas que están segmentadas en la segmentación se llama el token, y token de palabra significa esta unidad lingüística es palabra (MITKOV, 2003). Entonces, la clave de la segmentación automática de palabra en chino es sabe el ordenador qué es una palabra en el idioma de chino. Las **palabras** en chino son las unidades más pequeñas que tienen sentido para usar o decir solo(Zhu,1982). De hecho, esta definición es muy abstracta, no hay una descripción clara de qué es una palabra. Desde el punto de vista del ordenador, esta definición no explícita no se puede calcular. Para el ordenador, si no se puede calcular es equivalente a que no se puede operar.

Por ejemplo, la frase que hemos visto antes, “我喜欢吃冰淇淋。”, ya sabemos el significado de esta frase que es “A mí me gusta comer helado.”, segmenta en cuatro tokens, y además cada token está compuesto de diferente cantidad de caracteres. Pero ahora, si introduzco esta frase en el ordenador y le pido segmentar automáticamente, ¿puedes imaginar qué hará el ordenador con esta frase.?

Para segmentar palabras en chino, los humanos pueden usar el propio conocimiento para

[3] El Canon chino es un diccionario gratuito en línea con una gran capacidad de palabras, palabras, frases, modismos y otras formas de idioma chino(Traducción propia del chino a español)

determinar cuál es un token de palabra y cuál no y luego segmentarla. Pero para los ordenadores, trabaja automáticamente con estos, definitivamente opera con algoritmo. Qué hace el ordenador para comprender qué son los tokens de palabra en chino, este proceso de procesamiento es el **algoritmo** de la segmentación automática de palabras en chino(Wang y Guan, 2005). Y el algoritmo de la segmentación de chino lo explicaré al detalle en la parte final del trabajo.

5.4 Segmentación automática de palabras en chino

Por lo tanto, el proceso de segmentación del texto de chino en un conjunto de tokens de tipo palabras se denomina la **segmentación de palabra en chino**. Si el proceso lo hace el ordenador, se llama la **segmentación automática de palabras de chino** (Liu, 2000). El principio básico del método de segmentación automática de palabras de chino es analizar la cadena de entrada como el primer paso, luego segmenta las cadenas según los algoritmos, y al final, produce un conjunto de los tokens segmentados, cumpliendo la segmentación automática.
En la siguiente sección se revisarán los principales algoritmos de segmentación automática en palabras del chino para conocer cómo el ordenador puede realizar este proceso.

6. Algoritmos de la segmentación automática de palabras en chino

La sección anterior nos ha ofrecido una comprensión de la segmentación automática de palabras en chino. En esta sección, presentamos algunos algoritmos de segmentación automática que son más comunes actualmente, incluido sus principios y métodos de implementación. Estos algoritmos se pueden clasificar en tres categorías principales:

- algoritmos basados en diccionario,
- algoritmos basados en estadística,
- algoritmos basados en la comprensión del conocimiento.

6.1 Algoritmos basados en diccionario

Este tipo de algoritmo también se llama el método de segmentación mecánica. La premisa que utiliza este método es que el ordenador tiene que tener un diccionario de palabras segmentadas con el tamaño "grande y completo", luego el ordenador va a coger una cierta cantidad de

caracteres como una cadena desde la frase entrada y busca esta cadena en el diccionario. Una vez que la coincidencia es exitosa, significa que esta parte ya está segmentada, y al final si no coincide, esta cadena va a procesar con otro procesamiento (Liu, 2000). Este algoritmo tiene tres elementos básicos:

- ✧ El diccionario de palabras segmentadas,
- ✧ El orden por el cogido de las caracteres
- ✧ El principio de coincidencia.

El diccionario de palabras segmentadas tiene que ser lo más completo posible, con el desarrollo de la era, habrá muchas palabras nuevas que no existían antes, por lo tanto, este diccionario de palabras segmentada debería optimizarse continuamente. De acuerdo con el orden de lectura de los caracteres de una entrada textual, se divide en coincidencia hacia delante y coincidencia inversa. La coincidencia hacia delante significa procesar desde el primer carácter de la entrada, con lo contrario; la coincidencia inverso lee primero el último carácter de la entrada. De acuerdo con la situación de asignación de prioridad de diferentes longitudes de los caracteres , se puede dividir en la coincidencia máxima y la coincidencia mínima. La coincidencia máxima significa que el número del carácter de la cadena cogido por el ordenador es máxima, entonces el número total del resultado de los tokens de palabras de la entrada era más bajo. Y lo contrario, la coincidencia mínima significa el número del carácter de la cadena cogido es mínima, es decir el número del resultado de los tokens es más alta.

A modo de ilustración, se enumeran, a continuación, varios algoritmos que se usan de forma más común, tomando el algoritmo de la coincidencia máxima hacia delante como ejemplo.

6.1.1 El algoritmo de la coincidencia máxima hacia delante

De acuerdo con los tres elementos del algoritmo basado en diccionario que hemos mencionamos anteriormente, el algoritmo de la coincidencia máxima hacia delante procesa la entrada desde el primer carácter y además el número de los caracteres cogido el ordenador es máxima es decir la cantidad del resultado de los tokens es más pequeño.

Ejemplo:

“我是中华人民共和国的公民。[4] ”, el significado de esta frase es “Soy una ciudadana de la República Popular de China.”.

Sobre todo, según el principio de la coincidencia máxima, el resultado de la segmentación será con mínimo número de tokens de palabras. Si se hace una coincidencia con el diccionario, se puede segmentar con tal de coincidir en el diccionario, esta frase probablemente se segmenta como la segmentación opción 1 en el ejemplo 1 que está abajo, “我/是/中华/人民/共和国/的/公民。”, hay siete tokens de palabras. No podemos decir que está equivocado, porque todos los tokens de palabra se puede encontrar en el diccionario. Sin embargo, de acuerdo con el principio de la coincidencia máxima, este resultado es incorrecto, porque la cantidad del resultado de los tokens no es lo más pequeña posible. Para realizar la cantidad del resultado de los tokens de palabra en más pequeña, debería segmentar la palabra “中华人民共和国” como una palabra completa, es el nombre completo del país China. Se segmenta como el opción 2 en el ejemplo 1 “我/是/中华人民共和国/的/公民。”, que hay cinco tokens de palabra. La primera opción hay siete tokens, la segunda opción hay cinco tokens, de acuerdo con el principio de la coincidencia máxima, elige la segunda opción.

La frase entrada: 我是中华人民共和国的公民。

El significado de la frase: Soy una ciudadana de la República Popular de China.

La segmentación opción 1: 我/是/中华/人民/共和国/的/公民。

El significado de los tokens en español: yo/ soy/ china/ popular/ república/ de/ ciudadano

La segmentación opción 2: 我/是/中华人民共和国/的/公民。

El significado de los tokens en español: yo/ soy/ la República Popular de China/ de/ ciudadano

Ejemplo 1

Considerando un procesamiento hacia delante, esto significa el procesamiento de la entrada empieza por el primer carácter. Si usamos la frase anterior como el ejemplo, el primer carácter es “我”. Para evitar los posibles errores de la coincidencia máxima, usamos el método de reduce carácter para hacer la coincidencia. Suponemos que la palabra más larga de la frase entrada son

[4] El ejemplo se cita en el documento académicos *Investigación sobre técnicas clave de la segmentación de palabras en Chino*

ocho caracteres, así que tomamos los primeros ocho caracteres como una cadena y coincide con el diccionario. Si la coincidencia es exitosa, esta cadena se segmenta como un token, si no quitar un carácter de la cadena ha cogido, y coincide la nueva cadena con el diccionario, hasta al final la coincide con el diccionario. Después de la coincidencia de un token, toma otros los primeros ocho caracteres en la frase resta, y repite el ciclo anterior hasta que la última palabra está segmentada en token(Cao, 2009). El proceso de segmentación se presenta en la figura 1 que está abajo.

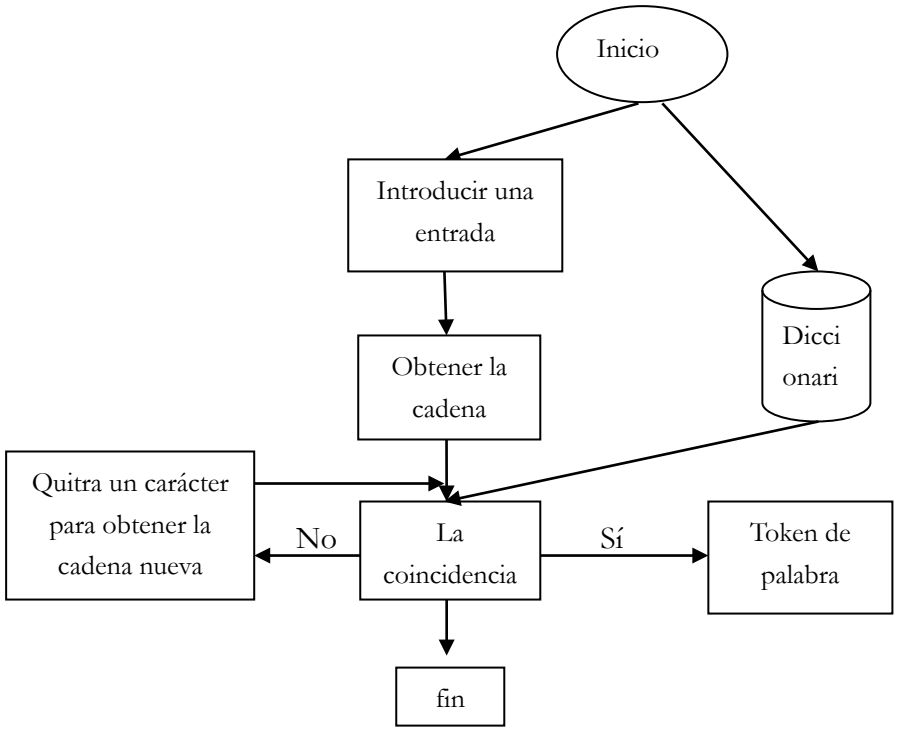

Figura 1[5]

6.1.2 El algoritmo de la coincidencia máxima inversa

El método y el proceso de la segmentación del algoritmo de la coincidencia máxima inversa es casi igual que el algoritmo de la coincidencia máxima hacia delante que hemos mencionado anteriormente, como se presenta en el figura 1. Pero la diferencia capital entre los dos algoritmos es el orden obtenido de la cadena. El procesamiento del algoritmo de la coincidencia hacia delante empieza por el primer carácter y termina en el último, pero en el algoritmo de la coincidencia inversa pasa lo contrario, el procesamiento empieza por el último carácter y termina

[5] La figura se cita en el documento académicos *Investigación sobre técnicas clave de la segmentación de palabras en Chino*

en el primero. Y además también hay otra diferencia en el procesamiento. En el caso de la coincidencia hacia delante, cuando la cadena elegida al final no coincide con el diccionario, se va a quitar el último carácter en la cadena. Pero en la coincidencia inversa, cuando no coincide la cadena, se va a eliminar el primer carácter en la cadena, porque la cadena se coge desde la parte posterior, por lo tanto se realiza así en la coincidencia inversa.

Los datos estadísticos muestran que la tasa de error de segmentar con el algoritmo de la coincidencia máxima hacia delante es 1/169, y además la tasa de error de usar la coincidencia máxima inversa es 1/245(Zhou, Zhu y Yang, 2001). Obviamente, el algoritmo de la coincidencia máxima inversa ha mejorado en la precisión comparado con el algoritmo de la coincidencia máxima hacia delante.

6.1.3 El algoritmo de la coincidencia máxima bidireccional

La coincidencia máxima como hemos visto, significa el resultado de la segmentación con menos tokens. Y además en este algoritmo no hay un requisito con la dirección para coger la cadena que va a coincidir con el diccionario, este algoritmo va a hacer una comparación entre el resultado de la segmentación de la coincidencia hacia delante y la coincidencia inversa. Este método es un algoritmo comúnmente usado en el sistema de la segmentación de palabras en chino, se aplica combinado con los algoritmos estadísticos, que puede reducir eficazmente la ambigüedad de la segmentación (Wang, 2014).

6.1.4 Ventajas y desventajas del algoritmo basado en diccionario

He enumerado tres algoritmos de la segmentación automática de palabras en chino que basa en el diccionario, ellos son la coincidencia máxima hacia delante, la coincidencia máxima inversa y la segmentación mínima. A través de los algoritmos anteriores, podemos ver fácilmente que los algoritmos basados en el diccionario son fáciles de cumplir., siempre que haya un diccionario suficientemente grande, las cadenas coinciden con el diccionario y las segmenta. Pero su defecto es que el nivel completo del diccionario va a determinar directamente la calidad de la segmentación, las palabras que no están en el diccionario no se pueden segmentar.

6.2 Algoritmos basados en estadística

A diferencia de los algoritmos basados en diccionario, los algoritmos basados en estadística no se basan en ninguna forma de diccionario, están completamente determinados por el cálculo de ordenador, y luego se segmenta.

Las palabras en chino se componen de caracteres múltiples o únicos. Por lo tanto, en el contexto, cuando más veces aparecen juntos los caracteres adyacente, es más probable que sea una palabra. La frecuencia o probabilidad de que un carácter u otro carácter sean adyacentes y aparezcan juntos se refleja directamente en el grado de credibilidad de una palabra (Long, Zhao y Tang, 2009). Cuando este valor es mayor que un número cierto, esta combinación de caracteres adyacentes es una palabra, y además se puede segmentar como un token. El proceso de este algoritmo es que primero corta las palabras, estas palabras como los artículos de recambios, luego el ordenador va a usar los modelos estadísticos para obtener los mejores resultados de la segmentación (Cao, 2009).

La segmentación automática de las palabras en chino basada en estadística se aplica a muchos modelos estadísticos. En general, para obtener el mejor resultado de la segmentación, el sistema de la segmentación automática combina varios modelos. Entre ellos, el modelos más utilizado es el modelo N-grama. En la siguiente subsección vamos a analizar este modelo.

6.2.1 El modelo N-grama

El modelo de N-grama, se basa en la suposición de que la enésima palabra está relacionada con la palabra n-1 y además no está relacionada con otras palabras. La probabilidad de que una oración completa sea correcta se calcula como el producto de la frecuencia de aparición de cada palabra que está en la frase dada la anterior (Wu, Wei y Li, 2001).

Desde un punto de vista estadístico, una oración en lenguaje natural puede estar compuesta de cualquier cadena de palabra, no hay las restricciones en al gramática. Pero, la probabilidad de la frase puede ser grande o pequeña, cuanto mayor es el producto de las probabilidades de las posibles secuencias de N palabras que contiene (Wu, Wei y Li, 2001). Por ejemplo, una oración O, ¿cómo calcular su probabilidad? Supongamos que la oración O está compuesta de una secuencia con n palabras P1, P2, P3, P4,...Pn. El proceso de calcular la probabilidad de la oración es el producto de la probabilidad de que aparezca la primera palabra P1 al principio multiplicado por la

probabilidad de la palabra P2 condicionada a que antes aparezca P1 y así sucesivamente:

$$P(O)=P(P1P2P3P4...Pn)$$

$$= P(P1)P(P2|P1)P(P3|P1P2)...P(Pn|P1P2...Pn-1)$$

Obviamente, este modelo es defectuoso. El espacio del parámetro es demasiado grande, especialmente cuando se toca la última palabra, tiene que ver con todas las palabras anteriores, incluso para el ordenador, también es un gran proceso. Para resolver este problema, el sistema de segmentación automática de palabras introduce la hipótesis de Markov. Vemos esta solución en el caso particular de los bigramas (n=2) o secuencias de dos palabras y trigramas (n=3) o secuencias de tres palabras en las dos subsecciones siguientes.

6.2.1.1 Bigrama

El hipótesis de Markov: La aparición de una palabra solo relacionada con una o varias palabras que están antes de ella(Wu, Wei y Li, 2001). Si la aparición de una palabra relaciona solamente con la palabra que está antes que ella, este método estadístico, lo llamamos **Bigrama**(Wu, Wei y Li, 2001). El algoritmo como se representa en la figura 3[6].

$$P(O) = P(W1W2W3...Wn)$$

$$= P(W1)P(W2|W1)P(W3|W2)...P(Wn|Wn-1)$$

Figura 3

La probabilidad de la palabra Wn venga después de la palabra Wn-1 en una frase es la cantidad de veces que aparece la dupla "Wn-1, Wn" en el corpus, divide por la vez en total de que aparece la palabra Wn-1. P(Wn|Wn-1) = count(Wn-1, Wn) /count(Wn-1).(Wu, Wei y Li, 2001) Vamos a ver un ejemplo en español, que sea más fácil para entender cómo se calcula.

[6] Este formula se cita en el libro *Un algoritmo de segmentación de palabras en chino basado en el modelo N-grama y el aprendizaje de máquinas*

Supongamos que nuestro corpus sea el ejemplo 2 [7] con tres frases.

1. <s>en un plato de trigo </s>
2. <s>tres tristes tigres </s>
3. <s>comen trigo </s>

Ejemplo 2

Los símbolos <s> y </s> son caracteres metalingüísticos que significan al principio y fin de la frase, para calcular la probabilidad de una palabra como el comienzo y el final de una frase. De acuerdo con nuestro corpus, podemos obtener los siguientes datos. P(en | <s>) = 1 / 3 = 0,33, P(un | en) = 1 / 1 = 1, P(plato | un) = 1 / 1 = 1, P(trigo | de) = 1/1 = 1, P(</s> | trigo) = 2/2 = 1, P(</s> | tigres) = 1/1 = 1, P(tres | <s>) = 1 / 3 = 0,33, P(de| plato) = 1/1 = 1. Los datos son las apariciones de bigrama en el corpus, solo enumera algunos datos que se usarán a continuación. Entonces, ¿Cómo deberíamos calcular la posibilidad de una frase? Por ejemplo la probabilidad de la frase "<s>en un plato de trigo</s> " es igual que P(en|<s>)* P(un|en)* P(plato|un)* P(de| plato)* P(trigo|de)* P(</s> | trigo), calcula con números la probabilidad de esta frase es 0.33*1*1*1*1*1 = 0.33. A partir del ejemplo, podemos encontrar fácilmente que la probabilidad de cada palabra solo tiene una relación con la palabra que está antes de ella, y no tiene nada que ver con ninguna otra palabra, así es el método de cálculo de bigrama.

La aplicación de bigrama en la segmentación automática de palabras en chino se refleja en el hecho de que después de cortar la frase en las posibles cadenas de caracteres, el algoritmo bigrama se usa para calcular la frase segmentada con más probable, para obtener el mejor resultado de la segmentación. Cuanto mayor es la probabilidad, más significativa es la frase segmentada. En el modelo de N-grama, además del bigrama, el más utilizado es el trigrama.

6.2.1.2 Trigrama

En los apartados anteriores, ya hemos visto sobre el algoritmo da bigrama, es decir, la aparición de una palabra solo está relacionada con la palabra anterior. Divide las palabras también incluidos los caracteres metalingüísticos en dos y dos como un grupo para calcular la probabilidad de la frase. Trigrama significa que la aparición de una palabra está relacionada con las dos palabras que

[7] Este ejemplo se cita en la pagina http://pdln.blogspot.com/2012/10/modelos-de-lenguaje-bigramas.html

está antes de ella(Wu, Wei y Li, 2001). Bigrama y trigrama son dos algoritmos más común en el modelo de N-grama.

6.2.2 Ventajas y desventajas del algoritmo basado en estadística

Como este algoritmo es puramente estadístico, algunas palabras que a menudo aparecen juntas en un sentido estadístico, pero en realidad no es una palabra. Por lo tanto, el resultado de la segmentación automática segmenta con el algoritmo que basada en estadística siempre lleva con problema. Entonces, el sistema de la segmentación de palabras chino rara vez usa el algoritmo basado con estadística, pero los modelos estadísticos se utilizan mucho para obtener el mejor resultado de la segmentación automática.

6.3 Algoritmos basados en reglas

Los algoritmos basados en reglas implementan la segmentación automática de las palabras a través del ordenador imitando la comprensión de los humanos con las frases del lenguaje natural. Su idea básica es realizar un análisis sobre el nivel sintáctico y el nivel semántico cuando hace la segmentación, utilizando información sintáctica e información semántica, y además segmenta las frases según el contexto (Zhang, 2005). Como en el ejemplo 3 [8], hay dos caracteres idénticos "和服", en estos dos ejemplo, los marco en negrita en los ejemplos. Los dos caracteres son totalmente iguales y aparecen juntos en estas dos frases, pero el resultado de la segmentación son diferentes, si lees estas dos frases, a través de la comprensión semántica, sería fácil para encontrar que en la primera frase, "和" y "服" se segmenta en dos tokens, y en la segunda frase, "和服" se segmenta en un token, este token significa el traje tradicional de Japón.

Frase 1: 制造业**和服**务业是两个不同的行业。

La segmentación de F 1º: 制造业/**和**/**服**务业/是/两个/不同的/行业。

El significado de los tokens de F 1º: la industria manufacturera/ **y**/ la industria de servicio/ son / dos/ diferentes/ industrias

El significado de F1º: La industria manufacturera y de servicio son dos industrias diferentes.

[8] El ejemplo se cita en el documento académicos *Investigación e implementación del sistema de segmentación automática de palabras en chino*

Frase 2: 日本的传统服饰是**和服**。

La segmentación de F 2º: 日本的/传统/服饰/是/**和服**。

El significado de los tokens de F 2º: japoneses/ Los trajes/ tradicionales/ son / kimonos.

El significado de F 2º: Los trajes tradicionales japoneses son kimonos.

Ejemplo 3

Suponemos que si el ordenador puede analizar y comprender los lenguajes naturales como los humanos, podría reducir muchos errores que se producen por que la máquina no entiende el semántico. Este método es más inteligente y concentra más atención a la comprensión de la máquina del lenguaje natural, por lo que también se llama el algoritmo basado en la comprensión. Este algoritmo por lo general, consta de tres partes:

- El subsistema de la segmentación de palabra
- El subsistema sintáctico y semántico
- La parte de control general.

Bajo de la coordinación de la parte de control general, el subsistema de la segmentación puede obtener el resultado de la segmentación posible, y luego el subsistema sintáctico y semántico juzga la frase segmentada a través de las informaciones sintácticos y semánticos, Este proceso imita el procesamiento de lenguaje natural de los humanos(Zhang, 2005).

Este algoritmo requiere una gran cantidad de conocimiento e información lingüística. Debido a la complejidad del conocimiento del lenguaje chino, es difícil organizar todas las informaciones de este idioma, por lo tanto, el sistema de la segmentación de palabras basada en la comprensión todavía se encuentra en una etapa experimental.

6.4 Aplicación de los algoritmos de segmentación en chino

El algoritmo es el núcleo de la segmentación automática de las palabras en chino. Un buen algoritmo de la segmentación juega un papel importante en el sistema de procesamiento de chino. Sin embargo, para cualquier sistema maduro, es imposible depender de un algoritmo para lograrlo, integra diferentes algoritmos.

Por ejemplo, Baidu, (el motor de búsqueda más usado en China), utiliza una combinación de algoritmos basado en el diccionario y la comprensión, debido a problemas con la velocidad,

Baidu no utiliza el algoritmo basado en estadística. (He y Wang, 2008)

Aunque tiene muchos algoritmos maduros, no significa que podemos resolver fácilmente las dificultades en el proceso de segmentación de palabras en chino. Como ya sabemos la complejidad del chino, es mucho más difícil para los ordenadores entender este idioma. Por ejemplo, los textos antiguos chinos son difíciles de entender para las personas, así que es casi imposible para el ordenador resolver las dificultades que entraña. También incluso tiene muchos problemas con el chino moderno.

7. Dificultad de la segmentación automática de palabras en chino

Las personas, cuando leemos un texto, podemos reconocer las palabras en el texto, a partir del conocimiento que tenemos sobre el lenguaje utilizado en ese texto. Las máquinas, igualmente, necesitan ese conocimiento lingüístico. El problema es cómo formalizarlo para que pueda programarse. En la actualidad, existen tres dificultades principales en la segmentación automática de palabras en chino (Zhang y Hao, 2005):

- La estandarización de la segmentación automática
- La segmentación automática de palabras ambiguas
- La identificación de nuevas palabras no registradas

7.1 Las dificultades en la estandarización de la segmentación automática

Como hemos visto antes, el chino es un lenguaje no segmentado muy típico, no tiene los espacios como el separador natural entre las palabras. Las **palabras** en chino son las unidades más pequeñas que tienen sentido para usar o decir solo(Zhu,1982). Esta definición es bastante abstracta, para las personas de diferentes niveles y con diferentes niveles educativos, el sentido del lenguaje y el nivel de comprensión de las palabras serán muy diferentes. Por lo tanto, desde el punto de vista estricto del cálculo, la segmentación automática de palabras en chino es un problema que no está claramente definido.

En la segmentación de palabras en chino, es necesario un diccionario, las palabras que existen en este diccionario deben segmentarse como un token, pero hasta ahora, no hay un diccionario autorizado con las palabras, por lo que todavía hay muchas dificultades en la segmentación entre

palabras y expresiones locuciones. En chino, hay algunos adjetivos y verbos que pueden producir estructuras deformadas. Por ejemplo en la Tabla de las deformaciones[9], hay cinco ejemplos, “打牌”, “高兴”, “有”, “看”,“相信” puede transformarse en “打打牌”, “高高兴兴”, “有没有”, “看一下”, “相不相信”.

Tabla de las deformaciones

El significado del prototipo	El prototipo en chino	El deformado en chino	El significado del deformado
Jugar a las cartas	打牌	打打牌	Jugar a las cartas
Contento	高兴	高高兴兴	Contento
Tener	有	有没有	Tener o no tener
Ver	看	看一下	Echar un vistazo
Creer	相信	相不相信	Creer o no creer

En la segmentación, estas deformaciones se segmentarán posiblemente en dos o más tokens, por ejemplo, “打打牌” se cortará en “打/打牌”. Sin embargo, todavía no existe un modelo operativo y razonable para la segmentación de las estructuras de deformación.

7.2 Las dificultades la segmentación automática de palabras ambiguas

La ambigüedad es un problema muy común en chino, por lo tanto, la segmentación de las palabras ambiguas es un problema difícil en la segmentación automática de palabras. Una oración con la misma forma se puede segmentar con diferentes resultados en diferentes casos o contextos, el significado de la oración que se expresará también va a cambiar. Hay dos ambigüedades principales en el proceso de segmentación automática de las palabras en chino, una es la ambigüedad de intersección y la otra es la ambigüedad combinada.

7.2.1 La ambigüedad de intersección

[9] Esta tabla se cita en el documento científico *Investigación e implementación del sistema de segmentación automática de palabras en chino(Cao,2009)*

La cadena de caracteres AJB tiene una ambigüedad cruzada, es decir, tanto AJ como JB son una palabra en chino, y el carácter J en este caso es una cadena de **intersección**(Sun y Liu, 2009). Como en el ejemplo 4 [10], la misma parte "的确定" en las dos frase, se puede segmenta en "的确/定", y también es posible segmentarla en "的/确定". Los tokens que están segmentada tienen su propio significado. De acuerdo con la definición de la ambigüedad de intersección, A es "的", J es "确", B es "定".

Frase 1: 这个计划**的确定**的不错。

La segmentación de F 1º: 这个/计划/**的确/定/**的/不错。

El significado de la segmentación: este/ plan/ realmente/ decide/ partícula/ bueno

El significado de F 1º: Este plan es realmente bueno.

Frase 2: 比赛**的确定**结果什么时候出来？

La segmentación de F 2º: 比赛/**的/确定/**结果/什么时候/出来？

El significado de F 2º: la competencia/ de/ determinado/ el resultado/ cuándo / sale

El significado de F 2º: ¿Cuándo sale el resultado determinado de la competencia?

Ejemplo 4

7.2.2 La ambigüedad combinada

Para una cadena AB, si se combina como una palabra, y además A y B en esta cadena también se pueden segmentar como palabras significativas. En este caso, se dice que la cadena AB tiene **la ambigüedad combinación** (Sun y Liu, 2009) Como el nombre completo de China que hemos mencionado en el Ejemplo 1, sería "中华人民共和国", entre ellos, "中华","人民" y "共和国" son tres palabras que llevan significado propios, y se combinan para formar otra palabra nueva. Para la ambigüedad combinación, podemos resolverlo mediante el algoritmo de la coincidencia máxima que hemos visto antes

7.2.3 La ambigüedad verdadera

Además de las dos ambigüedades presentadas anteriormente, también existe una ambigüedad

[10] El ejemplo se cita en el documento académicos *Investigación e implementación del sistema de segmentación automática de palabras en chino*

verdadera que significa una oración con ambigüedad, incluso las personas segmentan esta frase, sin el contexto, ni saben como hacerlo. Como en el ejemplo 5[11], una frase puede segmentarse de dos maneras, ambos resultados de la segmentación son sintácticamente correctos, pero la semántica es completamente diferente. Para obtener la correcta segmentación, debe hacer juicios basados en el contexto, por ejemplo, si en el texto se dice "Él está acostado en la cama.", en este caso, el primer resultado llevará con más posibilidad.

La frase: 他想起来了。

La segmentación posible: 他/想/起来了。

El significado de la frase segmentada: El/ quiere/ levantarse.

Otra segmentación posible: 他/想起来了。

El significado de la frase segmentada: El/ recuerda.

Ejemplo 5

Hay dos dificultades principales en la segmentación automática de palabras en chino, una es la ambigüedad que hemos mencionada anteriormente, y la otra es la identificación de palabras no registradas que presentamos en la siguiente subsección.

7.3 Las dificultades en la identificación de nuevas palabras no registradas

En el procesamiento de un texto ocurre que muchas de las palabras no están incluidas en el diccionario. Por ejemplo, los nombres de las personas, los nombres de lugares, las terminologías, las abreviaturas y muchas palabras nuevas que surgen durante el desarrollo y el cambio del idioma chino. Esta situación no se puede evitar porque no es posible registrar todas las palabras actuales en el diccionario.

Las palabras no registradas y la ambigüedad son los dos factores principales que afectan a la precisión de la segmentación de palabras en el idioma chino. Entre ellas, el impacto de las palabras no registradas es más grave. En el corpus real, los nombres propios y las terminologías representan una gran proporción, y estas no están en el diccionario en la mayoría de los casos. Las estadísticas muestran que la tasa de error causada por palabras no registradas es mayor que la tasa de error generada por las ambigüedades (Sun y Liu, 2009).

[11] https://wenku.baidu.com/view/f00cdd372cc58bd63086bd5d.html

8. La dirección de investigación futura de la segmentación en chino

La resolución de la ambigüedad y la identificación de las palabras no registradas siguen siendo las dos dificultades fundamentales para la segmentación de palabras en chino. Todos los algoritmos tratan de resolver estos dos problemas. En este sentido, en mi opinión, podemos pensar en dos direcciones de investigación futura de la segmentación en el idioma chino.

En primer lugar, el diseño de un algoritmo único que combine las posibles soluciones para la ambigüedad y para la identificación de las palabras no registradas. La mayor parte de los algoritmos revisados sólo pueden resolver un problema de los dos, y el no tratamiento del otro problema conduce a la reducción de la precisión del resultado de la segmentación.

En segundo lugar, utilizar soluciones híbridas que apliquen los diversos algoritmos de forma que se pueda mejorar la precisión en los resultados.

9. Conclusión

Este trabajo ha revisado el estado de la cuestión sobre la segmentación automática de palabras en chino, especialmente los algoritmos que se aplican y las dos dificultades fundamentales de la segmentación de palabras en chino.

De esta revisión se puede concluir que no existe un algoritmo que pueda resolver completamente todos los problemas encontrados en el proceso de la segmentación, y cada solución tiene su propia ventaja y desventaja.

El algoritmo basado en diccionarios puede segmentar con precisión todas las palabras que existen en el diccionario, pero puede haber ambigüedad. El algoritmo basado en estadísticas tiene un buen rendimiento en la identificación de la ambigüedad, pero la precisión no es tan buena como el algoritmo basado en el diccionario. La combinación orgánica de estos dos algoritmos puede resolver muchos problemas encontrados en la segmentación automática de palabras en chino.

Debido a la complejidad y variabilidad del idioma chino, todavía hay problemas que ninguno de los algoritmos puede resolver, como la ambigüedad verdadera. Por lo tanto, es necesario mejorar

los algoritmos de comprensión, de modo que el ordenador puede procesar como un cerebro humano. Todavía queda camino por recorrer en el estudio de la segmentación automática de palabras chino.

Bibliografía

MITKOV, R. (2003). *The Oxford handbook of Computational Linguistics*. Oxford University Press.

Yoav, G. (2017). *Neural Network Methods for Natural Language Processing*. A Publication in the Morgan & Claypool Publishers series.

Wang, X., & Guan, Y. (2005). *Procesamiento de lenguaje natural*. Beijing: Prensa de la Universidad de Tsinghua.31-44.

Zhu, D. (1982). *La gramática de chino*. Beijing: Prensa comercial.

Zhang, C., & Hao, T. (2005). La situación actual y las dificultades de la segmentación de palabras en chino. *Periódico de ciencia de simulación de sistemas*, *17*(1), 138–147.

Wu, Y., Wei, G., & Li, H. (2001). Un algoritmo de segmentación de palabras en chino basado en el modelo N-Gram y el aprendizaje automático. *Electrónica e informática*, *23*(11), 1148–1153.

Liu, K. (2000). *La segmentación automática de palabras en chino*. Beijing: Prensa comercial.

Cao, W. (2009). *Investigación sobre técnicas clave de la segmentación de palabras en Chino* (Tecnología informática). NUST, Nanjing.

Long, S., Zhao, Z., & Tang, H. (2009). Una descripción de los algoritmos de la segmentación de la palabra en chino. *Computer Knowledge and Technology*, *5*(10), 2605–2607.

Wang, F. (2014). *La confluencia de los algoritmos de la segmentación de palabras en chino*. Corpus Baidu (1–3).

Zhang, J. (2005). Algoritmo de la segmentación de las palabras en chino basado en reglas. *Informática y modernización*, (4), 18–20.

Sun, T., & Liu, Y. (2009). *Estado de la investigación y dificultades de la tecnología de la segmentación de palabras en chino*.

He, X., & Wang, W. (2008). Progreso de la investigación y aplicación de la tecnología de la segmentación de palabras en chino en recuperación de lenguaje natural. *Ciencia de la información*, *26*(5).

Zhou, C., Zhu, M., & Yang, Y. (2001). Investigación sobre algoritmo de segmentación de palabras en chino basado en Diccionario. *Ingeniería Informática y digital*, *37*(3), 68–71.