Procesamiento del Lenguaje Natural, Revista nº 44, marzo de 2010, pp 43-50 recibido 17-01-10 revisado 16-02-10 aceptado 05-03-10

Segmentación textual
Text segmentation

Fernando Chicharro Esteban
Universidad de Alcalá, Madrid
C/Río Ungría, 1. Portal 2, 3ºB.
19005 Guadalajara
ferchicharro@yahoo.es

Resumen: Una de las metodologías de investigación para identificar los mecanismos lingüísticos que representan la progresión temática es dividir en partes un texto o segmentarlo. Desde 1980, se han presentado dos clases de propuestas: los modelos probabilísticos y los jerárquicos, que se analizan y se comparan en este trabajo. También, se propone, a partir de los datos extraídos de los experimentos realizados, que los modelos probabilísticos, que emplean criterios de segmentación como la frecuencia y la distribución léxicas, no ofrecen resultados tan satisfactorios como los modelos jerárquicos, que ofrecen criterios para reconocer las intenciones comunicativas que un emisor puede desarrollar en un texto; existe un nivel de segmentación intraoracional (empaquetamiento informativo); y los modelos probabilísticos no ofrecen datos esclarecedores para identificar un segmento escrito en otra lengua.
Palabras clave: probabilístico, jerárquico, TextTiling, Focus stack, Centering.

Abstract: One of the methodologies to identify thematic continuity is to divide in parts a text or to segment it. From 1980, two classes of proposals have been proposed: the probabilistic models and the hierarchic ones, that are analyzed and compared in this paper. Also, one sets out, from the experiments, that the probabilistic models, that use the lexical frequency and the distribution, do not offer as satisfactory results as the hierarchic models, that offer criteria to recognize the communicative intentions; an intraorational level of segmentation (informative packing); and the probabilistic models do not offer enlightening data to language identification.
Key words: probabilistic, hierarchic, TextTiling, Focus stack, Centering

1 Planteamiento
1.1 Problema
Uno de los objetivos de los mecanismos lingüísticos para representar textualmente la progresión temática es la identificación y posterior segmentación de los subtemas que se desarrollan en un texto.

Según los criterios de análisis que se emplean para analizar la estructura de un texto, las principales propuestas pueden dividirse en dos grupos: los modelos que analizan la estructura del texto a partir de cálculos probabilísticos y los modelos que analizan la estructura jerárquica.

En este trabajo se comparan las propuestas más sugerentes y se constata que los modelos probabilísticos, que emplean criterios de segmentación como la frecuencia y la distribución léxicas, no ofrecen resultados tan satisfactorios como los modelos jerárquicos, que ofrecen criterios para reconocer las intenciones comunicativas que un emisor puede desarrollar en un texto.

Los datos extraídos de los experimentos también permiten plantear que existe un nivel se segmentación intraoracional (empaquetamiento informativo) que debería incorporarse como un tercer nivel de segmentación más exhaustivo a las segmentaciones globales y locales.

Además, si la segmentación efectuada se realiza a partir de cálculos probabilísticos se debería identificar la lengua que se utiliza en ese segmento. No obstante, los experimentos realizados con los modelos probabilísticos no ofrecen datos esclarecedores para identificar un fragmento escrito en otra lengua

1.2 Modelos probabilísticos
En los modelos probabilísticos se acepta que la estructura de un texto puede determinarse mediante cálculos probabilísticos. Algunos modelos son los siguientes:

ISSN 1135-5948

- (Morris y Hirst, 1991) se basa en la idea de que en cualquier texto existe unas cadenas léxicas que se corresponden con las intenciones comunicativas del emisor. La delimitación de estas cadenas permitirá entonces determinar cuáles son los segmentos de ese texto.
- En (Kozima, 1993) se propone un modelo basado en un indicador, denominado *Lexical Cohesion Profile (LCP),* que es un valor de la cohesión léxica entre palabras que permite determinar los límites de los segmentos. Se entiende la cohesión léxica como la similitud semántica entre palabras y se calcula mediante la activación de una red semántica que va construyéndose automáticamente a partir de un diccionario.
- En (Passonneau y Litman, 1993) se emplearon tres algoritmos basados en la distribución de los SNs (NP-A), de los sintagmas clave/marcadores discursivos (CUE-A) y de las pausas (PAUSE-A); intentaban replicar las segmentaciones manuales. Se calcularon los resultados con los algoritmos y en ningún caso se obtuvieron resultados mejores que en la segmentación manual; entre los algoritmos, los mejores resultados los ofrecía el NP-A y la combinación de NP-A y PAUSE-A.
- TextTiling fue presentado en (Hearst y Plaunt, 1993) y (Hearst, 1993, 1994a,b, 1997). Se compone de tres fases: división en palabras (tokenizar, lematizar, dividir el texto en pseudo-oraciones de un tamaño predefinido ($w \approx 20$) y guardar cada token analizado morfológicamente en una tabla con el número de la secuencia de tokens (*token-sequence*) en la que aparece y el número de veces que aparece en dicha secuencia de tokens); cálculo del valor de similitud léxica (calcular la similitud léxica de cada transición (*gap*) = *i* entre secuencias de tokens, calcular la puntuación léxica de la similitud entre los bloques mediante el producto escalar normalizado, representación gráfica y suavizar (*smoothing*) la gráfica); e identificación de límites (asignar un valor de profundidad (*depth score*), es decir, el valor de cada valle en la gráfica).

  Se comprobó que la agrupación por bloques ofrecía mejores resultados que, la introducción de vocabulario (Hearst, 1997) o las cadenas léxicas (Hearst, 1994b) y, por tanto, es la que se ha utilizado en la segunda fase del algoritmo.
- El modelo presentado en (Reynar, 1994) localiza los límites en el discurso a partir de la repetición léxica de ítems. Se utiliza el método gráfico de alineamiento de corpus bilingües *Dotplotting*.
- En (Litman y Passonneau, 1995) se presentaron dos métodos de segmentación: en el primero, el texto input se divide en bloques (*narratives*) que se consideran como una secuencia de límites posibles. En (Passonneau y Litman, 1993), se seleccionaba un límite como correcto cuando era clasificado como tal por cuatro o más de los siete sujetos experimentales. Sin embargo en (Litman y Passonneau, 1995), un límite era asignado como correcto si tres o más de los sujetos lo habían seleccionado. El segundo utiliza como input cada uno de los límites potenciales seleccionados con *Machine learning* C4.5, que genera árboles de decisión. Como output se obtiene un algoritmo de clasificación en forma de árbol de decisión que predice el tipo de límite potencial.
- En (Barzilay y Lapata, 2005) no se presenta un modelo de segmentación de textos, sino uno para calcular la coherencia textual, una vez que se ha segmentado con el modelo presentado en Barzilay y Lee (2004). Se basa en el modelo de coherencia local presentado en (Grosz, Joshi y Weinstein, 1983, 1986, 1995), denominado Centering. Así pues, en los textos que presentan coherencia local existen entidades cuya distribución tiene ciertas regularidades y, por tanto, la coherencia exige el cálculo del orden de aparición (*ranking*), asumiendo que es gradual.

1.3 Modelos jerárquicos

En los modelos jerárquicos se planteó el concepto de foco como centro de atención y un texto podía dividirse en segmentos, mostrando su estructura jerárquica, que se correspondían con las intenciones comunicativas del emisor. La relación entre estos segmentos originaba la coherencia global y dentro de un segmento, la coherencia local.

- Para explicar la coherencia global se propuso la teoría Focus stack (Grosz y Sidner, 1986, 1990). Se acepta que la estructura del texto está compuesta por tres componentes separados pero que interactúan

entre sí: la estructura lingüística (las oraciones que forman un texto y sus agrupaciones en segmentos); la estructura intencional (las relaciones jerárquicas que se establecen entre esos segmentos: dominio o precedencia); y la estructura del estado de atención (reflejo lingüístico de la evolución del estado de atención de un emisor).

Si, por una parte, se da una relación de dominio entre dos segmentos se debe quitar de la pila (*pop out from the stack*) el espacio focal correspondiente al segmento cuyas entidades estaban siendo prominentes, y debe ponerse encima de la pila (*push onto the stack*) el espacio focal correspondiente al nuevo segmento. Por otra parte, si se da una relación de satisfacción-precedencia entre dos segmentos, el espacio focal correspondiente al segundo se debe poner inmediatamente por encima del espacio focal correspondiente al primer segmento.

Este intercambio de espacios focales se conoce como intercambio de datos FIFO (First In First Out): el primer espacio focal que se haya puesto en la pila será el primero que se quite cuando se deba poner otro. Además, este intercambio se completa con el modelo FILO (First In Last Out): el primer espacio focal que se pone es el último que se quita.

Grosz y Sidner (1986: 192) propusieron que, además, debía utilizarse otra teoría que explicara los cambios de foco o centro de atención dentro de cada segmento. Esta teoría es la que conocemos como Centering.

- En (Grosz, Joshi y Weinstein, 1983, 1986, 1995) se propuso que la estructura jerárquica del discurso estaba formada por combinaciones de las intenciones comunicativas de los hablantes, su estado de atención y la forma de las expresiones referenciales que empleaban para explicar la coherencia local (dentro de un segmento).

  Se plantearon cuatro situaciones que reflejaban mayor o menor coherencia (ver Figura 1): o el Cb de una oración es el mismo que el de la oración anterior o no lo es; o el Cb de esa oración es igual que el Cp de su misma oración o no lo es.

	$Cb(U_i) = Cb(U_{i-1})$ o $Cb(U_{i-1}) = [?]$	$Cb(U_i) \neq Cb(U_{i-1})$
$Cb(U_i) = Cp(U_i)$	CONTINUA	CAMBIO SUAVE
$Cb(U_i) \neq Cp(U_i)$	RETENIDA	CAMBIO BRUSCO

Figura 1: Transiciones de Centering

Para gestionar estas situaciones, deben aplicarse dos reglas y tres restricciones:
Reglas:

- Si algún elemento de la Cf list (U_{i-1}) está representado con un pronombre en U_i, dicho elemento es el Cb (U_i).
- Continua > Retenida > Cambio suave > Cambio brusco.

Restricciones:

- Solo hay un Cb en cada enunciado.
- Cada elemento de la Cf list (U_i) tiene que estar representado en U_i.
- El Cb (U_i) es el elemento más importante de la Cf list (U_{i-1}), el cual está representado en U_i.

- En (Brennan, Friedman y Pollard (BFP), 1987) se presentó el primer algoritmo de Centering compuesto por tres fases: construir las posibles combinaciones (*anchors*) entre el Cb y la Cf list, <Cb-Cf> (crear un conjunto de expresiones referenciales (ERs), ordenar las ERs según su función gramatical, crear una lista de posibles centros prospectivos (Cf list), crear una lista de posibles centros retrospectivos (Cb's) y crear las combinaciones de <Cb-Cf> resultantes del producto de la cardinalidad de los dos pasos anteriores); filtrar las combinaciones <Cb-Cf> (filtro de contraíndices, filtro de restricción 3 y filtro de regla 1); y clasificar y ordenar (*rank*) las combinaciones (clasificar las combinaciones según las transiciones de Centering y ordenar las combinaciones según las transiciones de Centering).

2 Descripción de los experimentos

Se realizaron tres experimentos: uno para confrontar los resultados que ofrece TextTiling en la identificación de límites de subtemas con los que habían identificado unos sujetos experimentales en los mismos textos; otro experimento para comprobar si TextTiling puede identificar como límites, fragmentos de texto en otras lenguas; y otro, combinando Focus stack

más Centering para analizar los mismos textos y comparar los resultados obtenidos con la segmentación basada en criterios probabilísticos.

Además, como se muestra en los resultados de los experimentos, también se intenta constatar que los modelos probabilísticos pueden ofrecer resultados satisfactorios con textos no naturales.

2.1 Primer experimento: segmentación en subtemas

Se emplearon textos periodísticos monolingües (español peninsular): Noticia1-614 palabras, Noticia2-286 palabras, Noticia3-433 palabras, Noticia4-366 palabras.

La extensión media de cada oración ortográfica era ≈ 38 palabras, y cada párrafo ortográfico estaba compuesto por dos de estas oraciones, salvo alguna excepción en la que se reducía a una o se ampliaba a tres. Los signos de puntuación utilizados eran, exclusivamente, el punto (.) y la coma (,).

Con estos datos, se decidió modificar algunos parámetros de TextTiling: la extensión de cada secuencia de tokens (*w*) se fijó en 38, la cantidad de secuencias por bloque (*k*), en 2. Además, se manipularon esos parámetros incrementándolos y reduciéndolos progresivamente en una unidad para comprobar cuáles eran los que mejor se adaptaban a estos textos. Es decir, se comprobó la segmentación de cada texto modificando *k* desde 1 hasta 20 y *w* desde 38 hasta 28 y hasta 48. También se realizaron las segmentaciones de los textos con las modificaciones siguientes:

- k = 2/2, w = 38*2
- k = 2*2, w = 38/2
- k = 2/2, w = 38/2
- k = 2*2, w = 38*2
- k = 1, w = la extensión de cada texto en tokens
- k = la extensión de cada texto en tokens, w = 1

2.2 Segundo experimento: identificación de lenguas

Se emplearon siete textos bilingües y trilingües (catalán-inglés, catalán-español peninsular, español peninsular-francés-inglés y catalán-inglés-español peninsular):

- Textos académicos de 2º ciclo universitario: Texto 1: 191 palabras en catalán + 183 palabras en inglés = 374 palabras. Texto 2: 673 palabras en catalán + 14 palabras en español peninsular = 687 palabras. Texto 3: 1037 palabras en catalán + 60 palabras en inglés + 3223 palabras en español peninsular = 4320 palabras. Texto 4: 721 palabras en catalán + 105 palabras en inglés + 1099 palabras en español peninsular = 1925 palabras.
- Textos académicos de doctorado (tesis): Texto 5: 10458 palabras-Tesis redactada en español con citas en francés e inglés. Texto 6: 9909 palabras-Tesis redactada en español con citas en inglés. Texto 7: 9414 palabras-Tesis redactada en español con citas en inglés.

Se modificaron los parámetros de forma similar a la realizada en el experimento anterior. Se incrementó y se redujo en una unidad cada vez los parámetros *k* y *w*. Es decir, se comprobó la segmentación que realiza el algoritmo con el parámetro *k* fijado en 1 e incrementándolo en una unidad hasta 30; y el parámetro *w* desde 5 hasta 35.

Además, debido a que los textos académicos tienen unas características similares a los expositivos que analiza Hearst, se decidió emplear los parámetros por defecto (b = 2, n = 1, k = 6, w = 20) modificándolos de la siguiente manera:

- k = 6/2, w = 20*2
- k = 6*2, w = 20/2
- k = 6/2, w = 20/2
- k = 6*2, w = 20*2
- k = 1, w = la extensión de cada texto en tokens
- k = la extensión de cada texto en tokens, w = 1

Por último, se mantuvieron los parámetros por defecto del algoritmo para analizar un pseudo-texto compuesto por '…uno…dos…tres…cuatro…cinco…'.

2.3 Tercer experimento: Focus stack + Centering

Para desarrollar el análisis de los textos combinando Focus stack más Centering fue necesario solucionar dos inconvenientes: se debe segmentar previamente un texto para después aplicar Focus stack y, posteriormente, también es necesario dividir cada segmento en oraciones para poder ejecutar Centering. Para ello, se utilizaron los criterios de segmentación más

apropiados expuestos en (Walker 1998: 448-450).

3 Resultados de los experimentos

3.1 Primer experimento

Los textos periodísticos fueron segmentados por cinco sujetos para confrontar los resultados con la segmentación realizada por el algoritmo. Las similitudes no eran exactas debido, en parte, a que la única orientación que recibieron los sujetos era que debían indicar con un asterisco (*) dónde se cambiaba de subtema. Esto supuso que los textos presentasen una alta frecuencia de posibles segmentos (entre 10-11 para la primera noticia, entre 7-8 para la segunda, entre 9-10 para la tercera, y entre 7-8 para la cuarta); mientras que en las gráficas (ver Figura 2 y 3), el algoritmo sólo destaca dos o tres posibles segmentos.

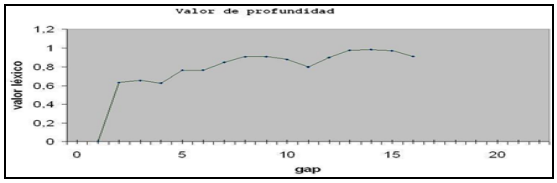

Figura 2: Segmentación Noticia 1

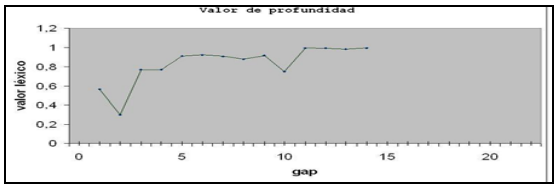

Figura 3: Segmentación Noticia 3

Además, se fijó el umbral conservador (HC, *High cutoff*), situando el límite en los valores de profundidad que excedían s-σ/2 y no se obtuvo ningún límite de segmento.

Todos estos datos hicieron pensar que el algoritmo de TextTiling debía ser probado previamente y con gran volumen de texto (tanto en número de textos como en su extensión) modificando los parámetros —casi de forma aleatoria— hasta localizar cuáles eran los que mejores resultados proporcionaban para un tipo de textos concreto.

Siguiendo a Hearst —y para intentar solventar esta segmentación tan confusa— se mantuvieron los parámetros que se emplean por defecto en el algoritmo para comprobar si los textos periodísticos también se segmentaban más convenientemente con las características de los textos expositivos. La segmentación obtenida era muy similar y no mejoró en ningún aspecto de la evaluación.

3.2 Segundo experimento

En los textos académicos bilingües y trilingües, la mayoría de los valores tenían un valor léxico = 1 o muy cercano (0.997, 0.998,…). En otras palabras, estos valores se interpretaron como que la similitud léxica era muy alta en esa parte de cada documento. No obstante, la revisión manual de los textos no coincidía plenamente con sus representaciones gráficas, debido a que, por ejemplo, en los fragmentos de tesis se utilizaban citas en otras lenguas con la suficiente extensión como para que fuesen detectadas por el algoritmo (ver Figura 4).

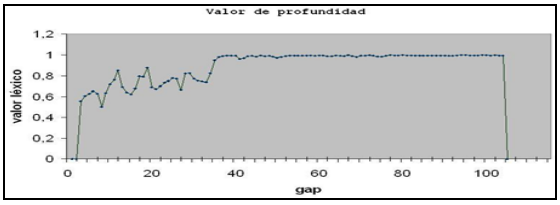

Figura 4: Segmentación tesis

En las prácticas redactadas por profesores universitarios y contestadas por alumnos, ocurría el mismo fenómeno que en el caso anterior, pero con la diferencia de que los fragmentos escritos en otras lenguas eran menos extensos. Esto hizo pensar que el algoritmo podría reconocer fragmentos de texto escritos en otras lenguas si se modificaban los parámetros de tal forma que la representación gráfica reflejase cualquier alteración en la similitud léxica, y fijando el umbral más liberal (LC (*Low cutoff*)) para que se seleccionasen más límites textuales. Para ello, se modificaron los parámetros como se ha indicado en la *Descripción de los experimentos*, pero los resultados fueron muy similares y no parece que podamos pensar que la segmentación que realiza el algoritmo de TextTiling sea completamente satisfactoria para la identificación de lenguas (ver Figura 5).

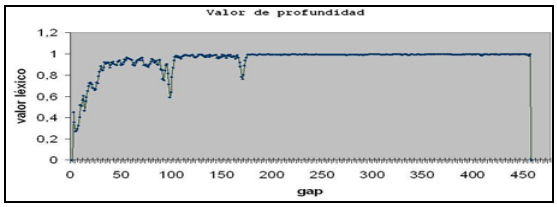

Figura 5: Segmentación práctica 2º ciclo universitario

Por el contrario, la gráfica del pseudo-texto muestra que el algoritmo identifica cada segmento con un valle (ver Figura 6 y 7). Este hecho corroboró la hipótesis de que los modelos probabilísticos pueden ofrecer resultados esclarecedores para la segmentación e identificación de lenguas, pero otorgando una gran importancia a la frecuencia y distribución léxicas y no al reconocimiento de la estructura informativa como los modelos jerárquicos.

…uno uno uno uno uno uno uno uno uno uno uno uno uno uno uno uno uno uno uno uno uno ... uno uno uno uno uno uno uno uno uno uno uno uno uno uno uno uno uno uno uno uno uno uno dos dos dos dos dos dos dos dos dos dos dos dos … dos dos dos dos dos dos dos dos dos dos dos dos dos dos tres tres tres tres tres tres tres tres tres tres tres tres tres tres tres tres tres tres tres tres tres tres tres tres … tres tres tres tres tres tres tres tres tres tres tres tres tres tres tres tres tres cuatro cuatro cuatro cuatro cuatro cuatro cuatro cuatro cuatro cuatro cuatro cuatro cuatro cuatro cuatro … cuatro cuatro cuatro cuatro cuatro cuatro cinco cinco cinco cinco cinco cinco cinco cinco cinco cinco cinco cinco… cinco cinco cinco cinco cinco cinco cinco cinco cinco…

Figura 6: Texto de lenguaje no natural (pseudo-texto)

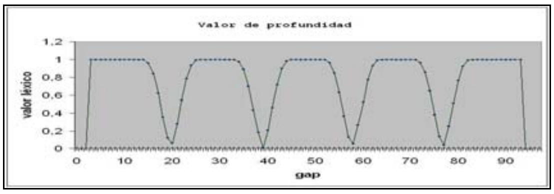

Figura 7: Representación gráfica de la segmentación del pseudo-texto

3.3 Tercer experimento

Para demostrar la viabilidad de combinar Focus stack y Centering, se presenta el análisis de un texto compuesto por lenguaje natural (novela) (ver Figura 8 y 9) junto con los resultados obtenidos del análisis del texto de la noticia.

Figura 8: Segmentación y relaciones de dominio y precedencia en una novela

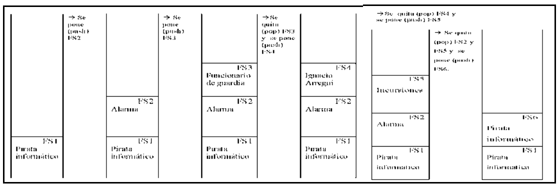

Figura 9: Intercambio FIFO y FILO en la pila de espacios focales

La pila de espacios focales se vacía cuando el texto finaliza completamente. Sin embargo, el texto corresponde a las dos primeras páginas de una novela, por lo que no ha terminado. No obstante, la pila se vaciará cuando se analice el siguiente segmento de la novela, debido a que se tratan otros objetos, otras propiedades y otras relaciones.

En cada segmento o espacio focal se aplica el algoritmo BFP'87 de Centering y los resultados obtenidos son los siguientes (ver Figura 10 y 11):

DS1	
U_i	No transición

DS2	
U_i	No transición
$U_i \rightarrow U_{i+1}$	Transición retenida
$U_{i+1} \rightarrow U_{i+2}$	Transición de cambio suave

DS3	
U_i	No transición
$U_i \rightarrow U_{i+1}$	Transición retenida
$U_{i+1} \rightarrow U_{i+2}$	Transición continua
U_{i+2}	Transición
$\rightarrow U_{i+3}$	continua

DS4	
U_i	No transición
$U_i \rightarrow U_{i+1}$	Transición continua
$U_{i+1} \rightarrow U_{i+2}$	Transición continua
$U_{i+2} \rightarrow U_{i+3}$	Transición continua
$U_{i+3} \rightarrow U_{i+4}$	Transición continua
$U_{i+4} \rightarrow U_{i+5}$	Transición continua
$U_{i+5} \rightarrow U_{i+6}$	Transición continua

DS5	
U_i	No transición
$U_i \rightarrow U_{i+1}$	Transición continua
$U_{i+1} \rightarrow U_{i+2}$	Transición continua

DS6	
U_i	No transición
$U_i \rightarrow U_{i+1}$	Transición continua
$U_{i+1} \rightarrow U_{i+2}$	Transición continua
$U_{i+2} \rightarrow U_{i+3}$	Transición retenida/continua

Figura 10: Transiciones de Centering en el texto de la novela

DS1	
U_i	No transición
$U_i \rightarrow U_{i+1}$	Transición continua/de cambio suave
$U_{i+1} \rightarrow U_{i+2}$	Transición retenida
$U_{i+2} \rightarrow U_{i+3}$	Transición continua

DS2	
U_i	No transición
$U_i \rightarrow$	Transición retenida
U_{i+1}	
$U_{i+1} \rightarrow U_{i+2}$	Transición de cambio suave

DS3	
U_i	No transición
$U_i \rightarrow U_{i+1}$	Transición retenida
$U_{i+1} \rightarrow U_{i+2}$	Transición de cambio suave

DS4	
U_i	No transición
$U_i \rightarrow U_{i+1}$	Transición continua

DS5	
U_i	No transición
$U_i \rightarrow U_{i+1}$	Transición retenida
$U_{i+1} \rightarrow U_{i+2}$	Transición continua

Figura 11: Transiciones de Centering en el texto de la noticia

La teoría de Focus stack y su componente Centering, no ofrecen buenos resultados cuando analizan pseudo-textos como el del experimento anterior: en primer lugar, el pseudo-texto se divide en segmentos (DS) que se denominan espacios focales (FS); en segundo lugar, se determinan las relaciones de dominio o precedencia. Como los segmentos no están compuestos por lenguaje natural, no se puede determinar si existe entre ellos una relación de dominio o precedencia. Por tanto, el análisis se interrumpe y no se continúa.

4 Conclusiones

A partir de los datos extraídos de los experimentos se concluye que los modelos probabilísticos segmentan los textos en bloques que no se corresponden exactamente con los subtemas que se desarrollan. Esto se debe a que el criterio de análisis de la frecuencia y la distribución léxicas adquiere demasiada importancia en los análisis; sin embargo, ofrecen resultados satisfactorios cuando se analizan pseudo-textos (lenguaje no natural).

Los modelos jerárquicos ofrecen resultados mejores que los probabilísticos cuando

segmentan un texto porque, entre otras cosas, reconocen la estructura intencional (Grosz y Sidner, 1986, Lochbaum, 1998) y no utilizan criterios como la frecuencia o la distribución léxicas.

De los datos obtenidos, también se ha constatado que existe un nivel de segmentación inferior que podría añadirse a la segmentación global de un texto y a la división local de cada segmento en oraciones (empaquetamiento informativo).

Además, el modelo probabilístico utilizado, TextTiling, no ofrece resultados esclarecedores para la identificación de un fragmento de texto escrito en otra lengua diferente a la que se utiliza en el texto.

Algunas cuestiones generales que deben tenerse en cuenta a la hora de realizar experimentos similares con el algoritmo de TextTiling son las siguientes:

- Los textos analizados sólo presentan caracteres occidentales, por lo que se hace necesario modificar el algoritmo para que reconozca, por ejemplo, alfabetos ideográficos.
- El algoritmo original sólo está diseñado para segmentar textos escritos en inglés, por lo que es necesario incrementar la *stop list* y el lematizador que se utiliza en el código fuente en la primera fase del algoritmo.
- Los parámetros que se utilizan para la segmentación dependen de las características de los textos que se analicen. Utilizar los parámetros por defecto del algoritmo para cualquier tipo de texto no ofrece buenos resultados.
- De la formulación inicial del modelo, se entiende que cualquier texto responde a una estructuración lineal entre los segmentos que lo componen. Sin embargo, es posible que un segmento esté vinculado temáticamente con otro que no es el inmediatamente precedente. Si aceptáramos ese orden lineal de la estructura textual, surgiría el mismo escollo que en el modelo jerárquico analizado, Centering (Poesio et al. 2004): hacen falta tres secuencias de tokens (dos enunciados en Centering) como mínimo para que pueda funcionar el algoritmo. Por un lado, Hearst advierte este problema pero sólo dice que son necesarias tres secuencias de tokens. Si, además, como estipula que —para los textos expositivos— la longitud de una secuencia de tokens es ≈ 20, la extensión mínima de un texto expositivo para ser analizado con TextTiling es 60 tokens. Por otro, en (Walker, 1998: 416) se propuso diferenciar entre los segmentos que están próximos linealmente y los que presentan una cercanía jerárquica o temática.
- Muchas gráficas —que representan muchos textos— pueden tener el mismo valor de profundidad (ver Figura 12).

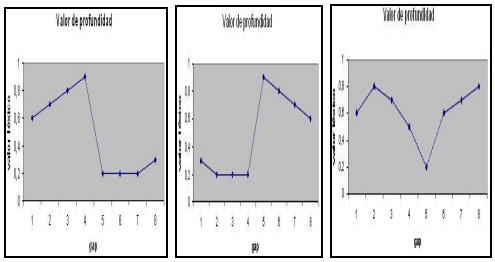

Figura 12: Diferentes gráficas, mismo valor de profundidad = 0.7

En el modelo combinado de Focus stack más Centering también hay inconvenientes que deben ser resueltos:

- Dividir en segmentos un texto para aplicar Focus stack.
- Dividir en oraciones cada segmento para ejecutar Centering.
- Dentro de una oración también hay relaciones informativas que deben analizarse para obtener una comunicación completa (empaquetamiento informativo).

Las perspectivas nuevas de segmentación apuntan hacia la idea más tentadora: combinar el enfoque probabilístico y el jerárquico (Barzilay y Lapata, 2005). No obstante, cualquier modelo que intente combinar estos enfoques debe tener en cuenta otra perspectiva para la segmentación de textos: la organización informativa del texto.

Dichas reglas deberían de ser las que organizan externamente un texto como los signos de puntuación (Figueras, 2001) o las reglas que normalicen el uso de unidades que codifican instrucciones de procesamiento como el contenido procedimental (Leonetti y Escandell, 2004: 1727): "los marcadores del discurso, las marcas de modalidad oracional, las partículas citativas y evidenciales, la entonación, los tiempos y modos verbales, los determinantes y pronombres definidos, los adverbios deícticos y focalizadores, y los mecanismos sintácticos que determinan la estructura informativa (por ejemplo, los que rigen las asignación del foco)".

Bibliografía

Barzilay, R. y M. Lapata. 2005. Modeling local coherence: An entity-based approach. En *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics*, páginas 141-148, Ann Arbor,

Barzilay, R. y L. Lee .2004. Catching the drift: Probabilistic content models, with applications to generation and summarization. En *Proceedings. of NAACL-HLT*.

Brennan, S., M. Friedman, y C. Pollard. 1987. A Centering approach to pronouns. En *Proceedings of the 25th Annual Meeting of the Association for Computational Linguistics*, páginas 155-162, Stanford, California.

Figueras, C. 2001. *Pragmática de la puntuación*. Octaedro. Barcelona.

Grosz, B. y C. Sidner. 1986. Attention, intentions and the structure of discourse. *Computational Linguistics*, 12(3):175-204.

Grosz, B. y C. Sidner. 1990. Plans for discourse. En Cohen, P., J. Morgan y M. Pollack, (eds.). *Intentions in communication*. páginas 417-444, MIT Press, Cambridge Massachusetts.

Grosz, B., A. Joshi, y S. Weinstein. 1983. Providing a unified account of definite noun phrases in discourse. En *Proceedings of the 21st Meeting of the Association for Computational Linguistics,* páginas 44-50, Cambridge, Massachusetts.

Grosz, B., A. Joshi, y S. Weinstein. 1986. Towards a computational theory of discourse interpretation. Versión final en Grosz, B., A. Joshi, y S. Weinstein. 1995.

Grosz, B., A. Joshi, y S. Weinstein. 1995. Centering: A framework for modeling the local coherence of discourse. *Computational Linguistics,* 21(2):255-265.

Hearst, M. 1993. TextTiling: A quantitative approach to discourse segmentation. Informe Técnico Sequoia 93/24, University of California. Computer Science Division.

Hearst, M. 1994a. Context and structure in automated full-text information access. Informe Técnico UCB/CSD-94/836 University of California. Computer Science Division.

Hearst, M. 1994b. Multi-paragraph segmentation of expository text**.** En *Proceedings of the 32nd Meeting of the Association for Computational Linguistics*, páginas 9-16.

Hearst, M. 1997. TextTiling: Segmenting text into multi-paragraph subtopic passages. *Computational Linguistics,* 23(1):33-64.

Hearst, M. y C. Plaunt. 1993. Subtopic structuring for full-length document access. En *SIGIR* **3, páginas 59-68.

Kozima, H. 1993. Text segmentation based on similarity between words. En *Proceedings of the 31st Annual Meeting of the Association for Computational Linguistics*, páginas 286-288, Columbus, Ohio.

Leonetti, M y M. V. Escandell, 2004. Semántica conceptual/ semántica procedimental. En M. Villayandre (ed.). *Actas IV Congreso de Lingüística General*, páginas 1727-1738, Arco/Libros, Madrid.

Litman, D. y R. Passonneau. 1995. Combining multiple knowledge sources for discourse segmentation. En *Proceedings of the 33rd Annual Meeting of the Association for Computational Linguistics*, páginas 108-115, Cambridge, Massachusetts.

Lochbaum, K. 1998. A collaborative planning model of intentional structure, *Computational Linguistics*, 24(4):525-572.

Morris, J. y G. Hirst. 1991. Lexical cohesion computed by thesaural relations as an indicator of the structure of text. *Computational Linguistics,* 17:21-48.

Passonneau, R. y D. Litman. 1993. Intention-based segmentation: human reliability and correlation with linguistic cues. En *Proceedings of the 31st Annual Meeting of the Association for Computational Linguistics*, páginas 148-155, Columbus, Ohio.

Poesio, M, R. Stevenson, B. Di Eugenio y J. Hitzeman. 2004. Centering: A parametric theory and its instantiations. *Computational Linguistics*, 30(3):309-363.

Reynar, J. 1994. An automatic method of finding topic boundaries. En *Proceedings of the 32nd Annual Meeting of the Association for Computational Linguistics*, páginas 331-332, Las Cruces, NM.

Walker, M. 1998. Centering, anaphora resolution and discourse structure. En Walker, M., A. Joshi, y E. Prince (eds.). *Centering theory in discourse*, páginas 401-435, Oxford, Clarendon Press.