Comparación de Sistemas para la Detección de Límites de Oraciones[3]

A Comparison of Sentence-Boundary Detection Systems

Celina Beltrán
Universidad Nacional de Rosario/ INDEC
Facultad de Ciencias Agrarias
Rosario, Argentina
beltranc@dat1.net.ar

Resumen

Se plantea la obtención de límites de oraciones (LO) mediante tres sistemas:

-Mx terminator: modela las decisiones que se toman al recorrer un texto y clasificar los elementos de S{., ?, !} en LO o no (sistema estadístico).

-SMORPH/MPS: distingue la separación de párrafos y de oraciones y la separación entre párrafos de la separación dada por nueva línea cuando no hay LO.

-XFST/Tokenize: agrega la detección de títulos a las funcionalidades del anterior.

Para evaluar los tres sistemas se utilizó un corpus de 277 oraciones.

Con el primero se obtiene 100% de cobertura y 96.9% de precisión en límite de oración (no discrimina final de párrafo y final de oración no final de párrafo). Con el segundo se obtiene 100% y 98,8% para precisión y cobertura de límites de párrafo, y 100% para precisión y cobertura de límites de oración que no finalizan párrafo. Con el tercero se detecta final de párrafo, final de oración no final de párrafo y títulos. Los valores de cobertura y precisión son 100% y 100% respectivamente para títulos, 100% y 100% para finales de párrafos y 100% y 99.5% para finales de oración no finales de párrafos.

Palabras claves: Límites de oraciones, Sistema estadístico, Sistema lingüístico, Comparación y evaluación.

Abstract

This paper explores sentence boundaries (SB) detection through three systems:

- Mx Terminator: it models the decisions made when running a text and classifying the elements of S{., ?, !} in SB or not (statistical system).

- SMORPH/MPS: it distinguishes between paragraph and sentence boundaries, as well as between paragraph and new-line boundaries when there is no SB.

- XFST/Tokenize: It adds title identification to the functions previously stated.

[3] Este trabajo pertenece a la tesis de Doctorado que realizo bajo la dirección del Dr. Gabriel G. Bés. Agradezco a François Trouilleux por sus imprescindibles sugerencias en la implantación en Xfst.

A corpus of 277 sentences was used in the evaluation of the three systems.

The first system rendered 100% coverage and 96.9% accuracy in sentence boundary identification (no discrimination between paragraph end and sentence-not paragraph end).

The second one rendered 100% accuracy and 98.8% coverage of paragraph boundaries, and 100% accuracy and coverage of boundaries of non-paragraph-ending sentences.

The third one identifies paragraph end, sentence-and-not-paragraph end, and titles. Its coverage and accuracy values are 100% and 100% respectively for titles, 100% and 100% for paragraph ending, and 100% and 99.5% for sentence-and-not-paragraph ending.

Keywords: Sentence boundaries, Paragraph boundaries, Title, Detection systems,

1. INTRODUCCION

Tradicionalmente, las tareas del procesamiendo de la lengua se han abordado desde dos perspectivas, una basada en información lingüística y la segunda basada en técnicas estadísticas.

El concepto de estadística que será utilizado es definido como sigue. Subyacente a un texto, existe un flujo de entidades que se caracteriza por sus frecuencias marginales y/o condicionales. Estas frecuencias se pueden extraer inductivamente mediante algoritmos apropiados aplicados a corpus de entrenamiento y, una vez extraídas, a partir de ellas, mediante cálculos adicionales que utilizan operaciones algebraicas se propone un sistema estadístico que se pretende válido para una categoría de textos. El sistema estadístico subyacente propuesto será aplicado a otros textos que el texto de entrenamiento, pero originados en la misma fuente. Al ser aplicados de esta manera, se deben obtener los resultados esperados: desambiguación de categorías morfo-sintácticas, desambiguación de interpretaciones semánticas, identificación de colocaciones, etc. Las funcionalidades del sistema estadístico subyacente están evidentemente condicionadas por las informaciones estadísticas extraídas inductivamente de los textos de entrenamiento, convertidas en los parámetros utilizados por el sistema.

En este trabajo se aborda la problemática de la obtención de límites de oraciones, esto es, la determinación del comienzo y del final de cada oración. Ésta es una etapa básica y clave del análisis automático de textos debido a que muchas tareas del procesamiento del lenguaje natural como la extracción y recuperación de información requieren un texto segmentado en oraciones.

En muchos casos, la identificación del final de una oración es una tarea simple, ya que un punto, un signo de exclamación o de interrogación a menudo señalan el límite de la oración. Sin embargo, existen muchos otros casos en los cuales el punto no está finalizando una oración, como puede ser un punto decimal en un número, una dirección de e-mail, una abreviación, o los signos de exclamación e interrogación que no necesariamente actúan siempre como límite de la oración.

Algunos de los problemas con los que uno se encuentra al segmentar un texto en oraciones son:

- Abreviaciones

- Siglas

- Direcciones de correo electrónico o páginas web

- Punto separador de miles o de decimales en los números

El problema a resolver es determinar si una ocurrencia de ".", "?" o "!" está marcando el final de una oración.

2. SISTEMAS UTILIZADOS Y EVALUADOS

Se aplican y se evalúan 3 sistemas para detectar los límites de oración (LO):

- Mx terminator: es un sistema estadístico que tiene por objetivo modelar las decisiones que se toman al recorrer un texto y clasificar los elementos de S{., ?, !} en LO o no.

- SMORPH/MPS: sistema implementado sobre Smorph que busca distinguir la separación de párrafos y de oraciones dentro de párrafos y además distinguir la separación entre párrafos de la separación dada por nueva línea cuando no hay LO.

- XFST/Tokenize: sistema implementado sobre la herramienta de XEROX, xfst/tokenize, donde además de las funcionalidades del anterior se le agrega la detección de títulos.

Para evaluar y comparar los tres sistemas, se utilizó un corpus de textos en español extraídos de las páginas web de periódicos argentinos. El mismo está formado por 277 oraciones y 7911 palabras. Estos textos fueron sometidos a cada una de estas propuestas y en base a los resultados obtenidos se calculó la cobertura y precisión alcanzada en cada aplicación.

3. MX TERMINATOR: HERRAMIENTA ESTADÍSTICA PARA LA DETERMINACIÓN DE LO

Muchos problemas concernientes al procesamiento del lenguaje natural pueden interpretarse como un problema de clasificación lingüística, en el cual el contexto es usado para predecir la clase lingüística. Los modelos de máxima entropía son usados para reformular el problema como un problema de clasificación estadística, en el que se estima la probabilidad de ocurrencia de la clase "A" con el contexto "B", esto es, p(A,B).

En el caso particular de detección del límite de oraciones, el objetivo es modelar las decisiones que se toman al separar un texto en oraciones, es decir, las decisiones que se toman al "recorrer" un texto y decidir si símbolos pertenecientes a S={. , ?, !}, son o no final de oración.

El modelo de máxima entropía asigna a cada elemento de S en el texto la probabilidad de ser final de oración, condicionado a su contexto.

Reynar J. y Ratnaparkhi A.[1] presentaron un modelo para identificar límites de oraciones basado en el enfoque de máxima entropía. Ellos consideraron cada una de las ocurrencias, en el texto de entrenamiento, que contiene al elemento de S. A la ocurrencia conteniendo el elemento de S se la llamó el "candidato" y se definió como prefijo y sufijo del candidato a la porción del mismo que estaba inmediatamente antes y luego del símbolo de S (contexto) con un blanco a la izquierda y a la derecha. Este modelo fue especificado para el inglés pero es posible adaptarlo a otros idiomas.

El modelo de máxima entropía utilizado expresa a la distribución de probabilidad conjunta

$$p(b,c) = \pi \prod_{j=1}^{k} \alpha_j^{f_j(b,c)}$$

donde $b \in$ {sí, no}, α_j son parámetros desconocidos del modelo y a cada parámetro le corresponde un f_j, el cual contiene la información contextual en forma de variable dicotómica. Por ejemplo, una característica podría ser:

$$f_j(b,c) = \begin{cases} 1 & si\ prefijo(c) = Mr\ y\ b = no \\ 0 & en\ caso\ contrario \end{cases}$$

Los parámetros de este modelo son estimados mediante el algoritmo *Generalized Iterative Scaling* (GIS). No obstante, estos parámetros estimados también son los que maximizan la verosimilitud ya que la solución por máxima entropía coincide con la de máxima verosimilitud. Este modelo estimado permite clasificar a un potencial límite de oración (candidato) de la siguiente manera:

$$\text{si} \quad p(sí / contexto) = \frac{p(sí, contexto)}{p(sí, contexto) + p(no, contexto)} > 0.5,$$

el candidato es clasificado como un límite de oración, en caso contrario, no.

El sistema MxTerminator fue desarrollado por Reynar y Ratnaparkhi utilizando el enfoque de máxima entropía para identificar oraciones en un documento. Fue entrenado sobre un corpus con 39441 oraciones de textos del Wall Street Journal y luego fue evaluado sobre otros dos corpus conformados por 20478 y 51672 oraciones respectivamente.

El sistema sólo requiere de un texto de entrenamiento el cual debe presentar una oración por línea, sin solicitar otra información. El sistema primeramente recorre el texto de entrenamiento y produce una lista de abreviaciones. Esta lista se construye considerando como abreviación a toda palabra del texto de entrenamiento que tiene un espacio antes y después de su ocurrencia y que finaliza con alguno de los símbolos ., ?, !, pero que no está indicado como final de oración. Por ejemplo, si el texto de entrenamiento fuese

> ***…….***
> ***El autor es A.L. Gómez.***
> ***Juan preguntó hay? y luego se fue.***
> ***Escribió "continuará… mañana" al terminar la página.***
> ***……..***

la lista contendría las siguientes tres ocurrencias

A.L.
hay?
"continuará…

Para la aplicación de este trabajo el texto de entrenamiento consta de 282 oraciones y 6657 palabras extraídas de textos periodísticos en internet de características similares al texto a analizar.

Luego define las funciones indicadoras que contendrán la información contextual. Retiene aquellos rasgos que se presentan con una frecuencia igual o superior a 10 en el texto de entrenamiento, es decir, sólo serán considerados por el sistema aquellos rasgos o características contextuales que se repitan 10 o más veces a lo largo del texto de entrenamiento. La adaptación al español sólo requiere de un texto de entrenamiento, del español, que presente una oración por línea.

Resultados obtenidos de la aplicación al español arrojaron una cobertura del 100% y una precisión de 96.9% (tabla 1).

Es importante destacar algunos aspectos del sistema basado en máxima entropía:

- Como se dijo anteriormente, los rasgos o características que se incorporan al modelo son aquellos que se presentan con una frecuencia igual o superior a 10 en el texto de entrenamiento. Esto significa que si, por ejemplo, el número con punto separador de miles, 10.000, se presenta una vez, el contexto prefijo=10 y sufijo=000 correspondiente a ese símbolo (.) no va a ser registrado para formar parte del modelo. Por lo tanto para que la información contextual de un punto correspondiente a una cifra con separador de miles sea considerada, ésta debe aparecer al menos 10 veces en el texto de entrenamiento.

- Las iniciales de nombres, como Pérez A.C., serán registradas en la lista de abreviaciones pero para que sean consideradas luego en los rasgos deben aparecer al menos 10 veces en el texto de entrenamiento.

- Cuando el sistema decide clasificar a un "candidato" como final de oración, va a colocar en el extremo de la oración a todo el candidato, por ejemplo, si decide que el punto de 10.000 corresponde a un final de oración, termina la oración luego del último 0, es decir

…….eran 10.000 casas.

es separado de la siguiente manera

…….eran 10.000
casas.

Esto también hace que cuando se omite el espacio al comenzar una nueva oración, como

La casa era linda. Era de Juan.

el sistema, reconoce las oraciones de la siguiente manera

Primera oración: ***La casa era linda.Era***
Segunda oración: ***de Juan.***

Por esta razón es que el sistema, si bien es posible que trabaje con un entrenamiento basado en un corpus de poco volumen, para que la precisión no disminuya, el texto de entrenamiento debe ser lo suficientemente extenso como para que el modelo pueda considerar la mayor cantidad de información contextual. Esto está relacionado con el tamaño de muestra en cualquier técnica de inferencia estadística. Si un evento es "raro" o poco frecuente, es necesario incrementar el tamaño de muestra para que se pueda estimar su probabilidad. En este sistema, el tamaño de muestra sobre el cual se estima el modelo corresponde a la cantidad de "candidatos" incluidos en el texto de entrenamiento.

Además, el desempeño de un modelo estadístico utilizado para clasificar, será mejorado aumentando el tamaño de la muestra (texto de entrenamiento), ya que los estimadores máximo verosímiles gozan de propiedades asintóticas (cuando el tamaño de muestra tiende a infinito). No obstante, si el tamaño de muestra es elevado, por ejemplo, 60.000 candidatos, considerar rasgos o características basadas en una frecuencia de 10 apariciones resultará en una gran cantidad de parámetros estimados.

Por último, en una técnica de inferencia estadística se desea que la muestra sea lo más "representativa" posible de la población en estudio, para que de esta manera el modelo refleje las características más sobresalientes de esta población. Esto sugiere que el texto de entrenamiento corresponda al mismo género que los textos sobre los cuales se va a aplicar el sistema. De esta manera estaría asegurado registrar las características pertinentes a los textos a analizar.

Tabla 1: Resultados obtenidos con Mx Terminator

	Existentes en el texto	**Identificados en MxTerminator**	**Coincidentes con los existentes**	**Evaluación por texto**	
				Cobertura	**Precisión**
Total límites*	277	286	277	**100,0**	**96,9**
Cant de palabras	7911				

*En este sistema no se discrimina el final de párrafo

4. SISTEMA IMPLEMENTADO SOBRE SMORPH Y MPS

4.1. Smorph y MPS

El software Smorph [2], analizador y generador, realiza la tokenización y el análisis morfológico, en una sola etapa y da como resultado las formas correspondientes a un lema (o a un subconjunto de lemas) con los valores correspondientes. Es una herramienta declarativa: la información utilizada por Smorph está separada de la maquinaria algorítmica, esto hace que se la pueda adaptar a distintos usos. Con el mismo software se puede tratar cualquier lengua sólo cambiando la información lingüística.

En Smorph deben declararse cinco tipos de información:

- Códigos Ascii
- Rasgos
- Terminaciones
- Modelos
- Entradas

El módulo post-smorph MPS [3] es un analizador, hace tratamientos anteriores al del resto de la sintaxis general de la oración con el objetivo de normalizar la entrada de la sintaxis y tener expresiones que satisfagan las mismas relaciones. En el trabajo con textos reales nos encontramos con expresiones que están fuera de una sintaxis estándar, por ejemplo: fechas, cantidades, todo lo que concierne a la sufijación y prefijación, incluido el tratamiento de clíticos, y las contracciones. MPS analiza estos microsistemas. Recibe en entrada una salida Smorph (en formato Prolog) y da a la salida otro formato según el analizador que se vaya a utilizar; MPS puede modificar las estructuras de datos recibidos en la entrada y ejecuta dos funciones principales: la Recomposición y la Correspondencia; la Recomposición a su vez puede ser de dos tipos diferentes, el Reagrupamiento y la División.

MPS es una herramienta declarativa, en la que, mediante reglas, se pueden expresar los valores de entrada (sobre dos o más estructuras de datos de la salida Smorph) y los valores de salida sobre la estructura reagrupada.

Las reglas que se declaren con la función de división provocan el efecto inverso. Son útiles para tratar las contracciones (por ejemplo. una ocurrencia de *del* en español), a fín de obtener en la salida una secuencia de entidades que sea análoga a las que Smorph asigna a las ocurrencias no contraídas en una cadena

Las reglas que se declaren con la función de correspondencia operan sobre una sola estructura de datos a la salida de Smorph y pueden modificarla en otra estructura de datos. Estas reglas permiten formular en Smorph descripciones básicas, generales, y adaptarlas después a la exigencia de cada analizador o de cada aplicación, o enriquecerlas con nuevos pares de <etiqueta=valor>.

Se busca distinguir la separación de oraciones dentro de los párrafos, de la separación de oraciones entre párrafos diferentes, y distinguir la separación entre párrafos, de la separación establecida por "nueva línea" cuando no hay límite de oración.

4.2. Sistema para detectar LO

El sistema propuesto en algunos casos fragmenta expresiones que no deben serlo (p.ej.las siglas), y en otros casos acepta expresiones que no están bien formadas (p.ej. '.).'), no obstante puesto que no aparecen en textos efectivos no genera inconvenientes.

Se busca reducir al máximo las entradas de Smorph. Prácticamente todas ellas están constituidas por caracteres que son signos de puntuación, con la excepción de las abreviaciones con '.' del tipo 'Sr.', las iniciales de nombre del tipo 'R.C.' o 'A.' y de las enumeraciones del tipo '1.', '2.'.

A la salida del análisis de Smorph, se suponen las siguientes reglas que deben aplicarse mediante MPS:

R1Si una ocurrencia es alguno de los casos:
'pfp' (punto final de párrafo)
'ifp' (interrogación final de párrafo)
'efp' (exclamación final de párrafo)

esa ocurrencia cierra una oración que es la oración final de un párrafo.

R2Si una ocurrencia es alguno de los casos:
'pf' (punto final)
'if' (interrogación final)
'ef' (exclamación final)

esa ocurrencia cierra una oración que NO es oración final de un párrafo.

El análisis de Smorph va a etiquetar algunas ocurrencias, como no siendo ni final de párrafo ni final de oración no final de párrafo. En el estado actual del sistema, las Reglas R1 y R2 se aplican aunque la misma ocurrencia reciba una etiqueta que comienza por 'n'. Por ejemplo, en

Juan lloraba. 10 maleantes lo atacaron

el '.' será 'pf' y 'npf' en el análisis Smorph, pero por R2 se opta por 'pf'.

En las entradas de Smorph se ha procurado especificar todos los '.', '?' y '!' que son o no, caracteres finales de oraciones, finales o no de párrafos.

Los ficheros Smorph tienen las siguientes características:

- En 'ascii' solo los /32 (blancos) están en ESPACES; /10 (nueva linea) y /9 (identación) están en SEPARATEURS pero no en ESPACES.

- Todos los caracteres de puntuación utilizados son SEPARATEURS.

- En las 'entrees' se ha utilizado una de las características principales de Smorph: prioridad al más largo. Así, por ejemplo la entrada '^xxx' tiene prioridad de aplicación a un texto de input sobre la entrada '^xx'

- Para anular los metacaracteres correspondientes a /10 y a /9 se utiliza el símbolo `. A continuación de este símbolo, se pulsa el símbolo correspondiente en el teclado, que no se imprime. Esta es la razón por la cual cuando se pulsa lo que corresponde a /10, es decir, se hace 'return', en una entrada, la entrada termina escribiéndose en dos líneas.

- De enumeraciones, abreviaciones izquierda tipo 'Sr.' y abreviaciones de nombres propios, sólo se incorporaron ejemplos.

Tipología de la utilización posible de '.':

- Expresiones con '.' en los extremos:
- Abreviación izquierda seguida de mayúscula: Sr., Ing. J. C. (iniciales de nombres)
- Abreviación izquierda que se supone en general no seguida de mayúscula: cm., kg.
- Abreviación derecha: .txt, .doc
- Enumeraciones: 1., 4.,
- Expresiones con '.' internos:

Siglas: C.N.B.A.
J.C. (iniciales de nombres sin blancos intermedios)
Cardinales: 15.000
Horas: 18.30
Siglas particulares: EE.UU
Direcciones mail: coca@patina.com.fr
URL : http://www.lsi.upc.es
Ausencia de centavos o de decimales: 10.- pesos

Se supone una tipología no idéntica pero análoga para '?' y para '!'.

En las entradas ('entrees'), se declaran:

- Abreviaciones izquierda que deben ser seguidas de mayúscula
- Esquemas de entradas combinadas con 'ALTERS' para iniciales de nombres. Ejemplo:
- esquema N._N. para J. C. -Juan Carlos y N._. Para inicial simple
- Enumeraciones de '1.' a '10.'
- Los '.' internos se tratan con ^.^
- Los puntos finales de siglas y de abreviaciones de tipo 'cm.' se tratan con '^.|' y con '^.' (Recordar que '|' exige que a la derecha de '.' no haya minúscula). Esta solución va a dar

ambigüedad cuando el '.' está seguido de cardinales. En el sistema actual se resuelve siempre el '.' como límite cuando hay ambigüedad.

El resto de las entradas se organizan alrededor de las características generales anteriores. Se trata de eliminar al máximo los '.' que no son límites, aprovechando las secuencias con ',', ';', etc.

Es importante destacar que sólo será necesario modificar las entradas correspondientes a las abreviaciones cuando se desee adaptar el sistema a cualquier otro idioma.

Sobre el corpus de evaluación fue aplicado y se obtuvieron los resultados de la tabla 2. La cobertura general fue del 99.6% y la precisión del 100%.

Tabla 2: resultados obtenidos con Smorph y MPS

	Existentes en el texto		**Etiquetado en smorph/mps**		**Coincidentes con los existentes**		**Evaluación por texto**	
	pfp	**pf**	**pfp**	**pf**	**pfp**	**pf**	**Cobertura**	**Precisión**
Límites finales de párrafo	81		80		80		**98,8**	**100,00**
Límites finales de oración		196		196		196	**100,0**	**100,00**
Total límites	277		276		276		**99,6**	**100,00**
Cant de palabras	7911							

5. SISTEMA IMPLEMENTADO SOBRE XFST

Xerox finite-state tolls (XFST) es una interface interactiva que provee acceso a algoritmos básicos de cálculo de estados finitos [4]. Además, provee un compilador para extender metalenguajes de expresiones regulares. Este compilador posee una regla de reemplazo muy útil, particularmente en esta aplicación.

En este trabajo, además de la herramienta XFST se utiliza otra llamada TOKENIZE. Tokenize es una aplicación que ejecuta un network de estados finitos a un texto de entrada y lo divide en "tokens". Esta aplicación es usada para que en un texto, que se desea segmentar en oraciones, le coloque "balizas" a cada final de oración determinado por ".", "?" o "!".

A continuación se detallan las distintas funciones definidas en XFST

FST1 - Introducen los límites finales de párrafo:

<Xfp/>, con X en {s, p, i, e, a}

a la derecha de '...', '.', '?', '!', ':' cuando éstos no tienen como contexto derecho ningún o varios blancos y una nueva línea. El orden de precedencia de FST1 en relación a FST1A es significativo.

FST2 - Introducen los límites finales de párrafo:

<Xfp/>, con X en {s, p, i, e}

a la derecha de ')' ou de '"' cuando éstos tienen como contexto izquierdo '...', '.', '?', '!', opcionalmente seguidos de un blanco y como contexto derecho ningún o varios blancos y una nueva línea. El orden de precedencia de FST2 en relación a FST2A es significativo.

FST3 - Introducen los limites finales de oración no finales de párrafo:

<Xf/>, con X en {s, p, i, e}

cuando éstos tienen como contexto derecho uno o varios blancos seguidos de una mayúscula, o dígito o '(' o '"'. El orden de precedencia de FST3 en relación a FST3A es significativo.

FST4 - Introducen los limites finales de oración no finales de párrafo:

<Xf/>, con X en {s, p, i, e}

a la derecha de ')' ou de '"' cuando éstos tienen como contexto izquierdo '...', '.', '?', '!', opcionalmente seguidos de un blanco y como contexto derecho uno o varios blancos seguidos de una mayúscula, o dígito o '(' o '"'. El orden de precedencia de FST4 en relación a FST4A es significativo.

FST5 - Elimina los '</pf>', que han sido introducidos por FST3, cuando están precedidos por una abreviación de tipo 'Cf., cf., Sr., Ing.', o de tipo 'J.', o para el español 'J. C.' o 'J.C.' o para el francés 'J-C.' (abreviaciones de iniciales de nombres propios), o para las enumeraciones, previstas de '1.' a '9.'. De este modo, en todos los casos siguientes con la excepción de los '.' finales de oración o de párrafo, no se tiene baliza después de '.'. La definición de ABREV es determinante para el funcionamiento de FST5.

FST6 - Elimina los '</pf>', que han sido introducidos por FST3, cuando están precedidos por una enumeración de sección de un artículo científico. El orden de precedencia de FST3 en relación a FST6 es significativo.

FST7 - Eliminan respectivamente los '</sf>, </pf>, </if>, </ef>,' que han sido introducidos por FST3, cuando están precedidos por los símbolos que los han introducido y están seguidos de un blanco, de '"' et de la baliza correspondiente de fin de párrafo. El orden de precedencia de FST2 y de FST3 en relación a FST5 es significativo.

TIT1 / TIT2 - Introducen a la izquierda y a la derecha de los títulos las balizas :

<tit> </tit>

Los títulos pueden estar precedidos o no por los dígitos que los enumeran, pero, en todos los casos, están precedidos y seguidos por dos o más nuevas líneas, no tienen una baliza interna y no terminan ni por '.' ni por ':'. TIT1 asocia las balizas a los títulos precedidos por dígitos y TIT2 a los que no están precedidos por dígitos.

A diferencia de los dos sistemas anteriores en éste es posible detectar los títulos, además del final de oración y de párrafo. Los resultados hallados en el corpus de evaluación son los presentados en la tabla 3.

Evaluando la detección de títulos y finales de párrafos se obtuvo una cobertura y precisión del 100% en ambos casos mientras que para los límites de oraciones que no son finales de párrafos estas cifras fueron de 100% y 99.5% respectivamente.

Tabla 3: resultados obtenidos sobre XFST

		Etiquetado en AUTOMATAS		**Coincidentes con los existentes**		**Evaluación por texto**	
	Existentes	**pfp**	**pf**	**pfp**	**pf**	**Cobertura**	**Precisión**
TITULOS	26	26		26		**100,00**	**100,00**
Puntos finales de párrafo	81	81		81		**100,00**	**100,00**
Puntos finales de oración	196		197		196	**100,00**	**99,49**
Cant de palabras	7911						

6. DISCUSIÓN

Con el sistema estadístico se obtiene un 100% para la cobertura y un 96.9% de precisión para el límite de oración, sin poder discriminar entre final de párrafo y final de oración no final de párrafo.

Este desempeño es superado por el sistema sobre Smorph/MPS en el que se obtienen los valores de 100% y 98,8% para precisión y cobertura de límites de párrafo, y 100% para precisión y cobertura de límites de oración que no son límites de párrafo, adicionando la discriminación del final de párrafo.

Con el sistema basado en la herramienta de Xerox, Xfst, se agrega una nueva funcionalidad detectando título además de final de párrafo y final de oración no final de párrafo. Los valores de cobertura y precisión son 100% y 100% respectivamente para títulos, 100% y 100% para finales de párrafos y 100% y 99.5% para finales de oración que no son finales de párrafos.

Si bien para los tres sistemas los valores de precisión y cobertura son elevados, es importante observar las frecuencias absolutas de los errores y aciertos de cada sistema. La tabla 4 evidencia que el sistema basado en el modelo de máxima entropía clasificó a toda ocurrencia {. , ?, !} como límite de oración.

El porcentaje de clasificación correcta,

$$p = \frac{\textit{número de candidatos bien clasificados}}{\textit{número total de candidatos}}.100$$

,

proveniente del sistema estadístico es 96.8%, mientras que el correspondiente al sistema bajo SMORPH y XFST es 99.6%. Sin embargo, si bien los tres porcentajes son satisfactorios, cuando se construyen los intervalos de confianza del 95% para el porcentaje de clasificación correcta bajo cada uno de los sistemas se encuentra que el hallado con el modelo de máxima entropía se encuentra por debajo de los otros dos (tabla 5).

Tabla 4: Frecuencias absolutas obtenidas en la clasificación de límite de oración.

Mx TERMINATOR **Sistema**

		Sí	**No**	**Total**
Reales	**Sí**	**277**	0	277
	No	9	**0**	9
	Total	286	0	286

SMORPH/MPS **Sistema**

		Sí	**No**	**Total**
Reales	**Sí**	**276**	1	277
	No	0	**9**	9
	Total	276	10	286

XFST/TOKENIZE **Sistema**

		Sí	**No**	**Total**
Reales	**Sí**	**277**	0	277
	No	1	**8**	9
	Total	278	8	286

Tabla 5: Intervalos de confianza del 95% para el porcentaje de clasificación correcta.

Sistema	Porcentaje de clasificacion correcta	Límite inferior del intervalo (95% de confianza)	Límite superior del intervalo (95% de confianza)
Mx Terminator	96.8	94.7	98.8
SMORPH	99.6	98.9	1.00
XFST	99.6	98.9	1.00

Referencias

[1] Reynar J. and Ratnaparkhi A. A maximum entropy approach to identifying sentence boundaries. In: Proceedings of Fifth Conference on Applied Natural Language Processing. 1997

[2] Aït-Mokhtar, Salah L'analyse présyntaxique en une seule étape. Tesis doctoral. Universidad Blaise-Pascal/GRIL, Clermont-Ferrand, 1998.

[3] MPS ha sido especificado en el GRIL por Caroline Hagège, José Rodrigo, Gabriel G. Bès y Faiza Abacci, e implantado en C++ en un contexto Windows por Faiza Abacci.

[4] Beesley K.R. and Karttunen L. Finite State Morphology. CSLI Publications, Stanford University. 2003.