Comparación y evaluación de dos etiquetadores[1]

Comparison and Evaluation of Two Labelers

Celina Beltrán

Universidad Nacional de Rosario
Facultad de Ciencias Agrarias
Rosario, Argentina
beltranc@dat1.net.ar

Abstract

This paper evaluates two labelers:

- a system that uses bigram models for lexical categories, enriched with information from some words from the lexicon in specific contexts.
- a system based on the use of the tools SMORPH and MPS to obtain the morphological analysis of Spanish texts as well as the resolution of ambiguities through the application of rules defined in MPS.

In order to evaluate and compare both systems, a corpus of Spanish texts from Argentinean newspaper web pages was used. The application obtained 96.36 % accuracy for the statistical system and 99.17 % accuracy for the SMORPH/MPS-based system. Reliability intervals are (95.95%; 96.77%) and (98.97%; 99.37%), respectively.

Keywords: Labeler, Markov Model, Bigram, Linguistic system, Comparison and evaluation.

Resumen

En este trabajo se evalúan dos etiquetadores:

- sistema que utiliza modelos de bigramas de categorías léxicas enriquecidos con información de ciertas palabras del vocabulario en determinados contextos.
- sistema basado en la utilización de las herramientas SMORPH y MPS para obtener el análisis morfológico de textos en español como así también la resolución de ambiguedades mediante la aplicación de reglas definidas en MPS.

Para evaluar y comparar los dos sistemas, se utilizó un corpus de textos en español extraídos de las páginas web de periódicos argentinos. En la aplicación se obtuvo una precisión del 96.36% para el sistema estadístico y del 99.17% para el sistema basado en SMORPH/MPS. Los intervalos de confianza del 95% son respectivamente (95.95% ; 96.77%) y (98.97% ; 99.37%).

Palabras claves: Etiquetador, Modelo de Markov, Bigrama, Sistema lingüístico, Comparación y evaluación.

[1] Este trabajo pertenece a la tesis de Doctorado que realizo bajo la dirección del Dr. Gabriel G. Bès.

1. INTRODUCCION

En este trabajo se evalúan dos etiquetadores, uno basado en modelos de Markov y otro que utiliza información lingüística.

El etiquetador estadístico es el presentado en Pla (2001) [1]. Es un sistema de etiquetado morfosintáctico que utiliza modelos de bigramas de categorías léxicas enriquecidos con información de ciertas palabras del vocabulario en determinados contextos. Este sistema ha sido evaluado experimentalmente sobre el corpus en castellano LexEsp. Los autores aseveran que estos modelos, a los que llaman "Modelos contextuales especializados", mejoran el desempeño para resolver ciertas ambigüedades.

El segundo sistema está basado en la utilización de las herramientas SMORPH y MPS para obtener el análisis morfológico de textos en español como así también la resolución de ambiguedades mediante la aplicación de reglas definidas en MPS.

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.

2. SISTEMA BASADO EN MODELOS DE MARKOV

2.1 Utilización de modelos de Markov para el etiquetado

Sea el conjunto de categorías léxicas $C = \{c_1, c_2, \cdots, c_N\}$ y el vocabulario correspondiente a la aplicación. El objetivo del modelo es hallar, para una frase de entrada $w = w_1, \cdots, w_T$, la secuencia de categorías de máxima probabilidad, esto es, hallar la secuencia de categorías o etiquetas $\hat{\mathbf{c}} = (c_{(1)}, c_{(2)}, ..., c_{(T)})$ de modo tal que sea máxima su probabilidad dado que se ha observado la frase de entrada w. Si llamamos con $P(\mathbf{c}/\mathbf{w})$ a la probabilidad de la secuencia de etiquetas $\mathbf{c}$ condicionada a la frase $\mathbf{w}$, la solución brindada por un modelo de Markov es la secuencia de etiquetas $\hat{\mathbf{c}}$ tal que la probabilidad $P(\hat{\mathbf{c}}/\mathbf{w})$ es máxima, esto es la solución a la ecuación

$$\hat{c} = \arg\max_{c \in C^T} P(\mathbf{c}/\mathbf{w})$$

la cual, aplicando la definición de probabilidad condicionada y la regla del producto para la probabilidad conjunta de dos eventos, puede expresarse como

$$\hat{c} = \arg\max_{c \in C^T}\left(\frac{P(\mathbf{c},\mathbf{w})}{P(\mathbf{w})}\right) = \arg\max_{c \in C^T}\left(\frac{P(\mathbf{c})P(\mathbf{w}/\mathbf{c})}{P(\mathbf{w})}\right). \qquad (2.1.1)$$

Para hallar la solución a la ecuación (2.1.1) sólo es necesario maximizar el numerador adicionando ciertos supuestos. Un modelo de Markov de primer orden o bigrama, asume que la probabilidad de una categoría $c_{(k)}$ en la posición (k) de la frase de entrada sólo depende de la categoría asignada en la posición (k-1), independientemente de la posición en la secuencia (no depende de k). Por esta razón, el problema de hallar la solución a (2.1.1) se reduce a resolver la siguiente ecuación

$$\arg\max_{c_{(1)}\ldots c_{(T)}}\left(\prod_{i=1}^{T} P(c_{(i)}/c_{(i-1)})P(w_{(i)}/c_{(i)})\right). \qquad (2.1.2)$$

En este modelo de Markov, las categorías léxicas son los estados, las probabilidades contextuales son las probabilidades de transición entre estados y las probabilidades léxicas son las probabilidades de emisión de símbolos desde cada estado. El etiquetado se lleva a cabo mediante el algoritmo de Viterbi.

2.2. Especialización del modelo

Este sistema introduce información para ampliar el contexto considerado mediante la incorporación de ciertas palabras y categorías léxicas. De esta manera, se pueden establecer ciertas restricciones de contexto ligadas al léxico.

Los criterios para definir las palabras que van a especializarse pueden ser:

- Las palabras más frecuentes
- Las palabras con mayor error de etiquetado
- Las palabras pertenecientes a categorías cerradas.

Así, de alguna manera, se logra introducir conocimiento lingüístico a los modelos.

La especialización consiste en la definición de una función que permite el reetiquetado del texto de entrenamiento ampliando el conjunto de etiquetas. Como ejemplo de especialización, presentado en Pla (2001), se puede considerar la modelización en forma separada de la ocurrencia de la palabra *que* como pronombre relativo, de la aparición como conjunción subordinante. Para llevar a cabo esto es necesario incluir en el conjunto W_{esp} la palabra *que*. Para este caso la función de especialización reemplaza, en el corpus de entrenamiento, el par de <palabra, etiqueta> <que,PR> por el par <que, (que PR)> y el par <que,CS> por el par <que, (que CS)>. De esta manera aparecen dos nuevos estados: (que CS) y (que PR) que amplían el conjunto original de etiquetas.

2.3. Etapas del Sistema

El sistema consta de dos etapas: el aprendizaje sobre el conjunto de entrenamiento y el etiquetado de un nuevo texto.

En la **etapa** de **aprendizaje o entrenamiento** se utiliza un corpus de entrenamiento especializado (etiquetado de la manera previamente descripta). El conjunto de etiquetas que se utiliza es el conjunto de 65 categorías *PAROLE*.

Sobre ese corpus etiquetado se estiman o “aprenden” los modelos de Markov (bigramas) por máxima verosimilitud. Las probabilidades se obtienen a partir de las frecuencias de ocurrencia de las palabras, categorías y de la ocurrencia de cada palabra en cada categoría.

Para las palabras desconocidas se asume que pueden pertenecer a cualquier categoría y la misma es aproximada mediante el cálculo de las probabilidades correspondientes.

Una vez que el sistema fue entrenado, esto es, se tienen estimadas todas las probabilidades de transición entre categorías y las probabilidades de cada palabra en cada categoría, se procede a utilizar dichas probabilidades para etiquetar un nuevo texto (etapa de etiquetado).

El sistema de etiquetado tiene por entrada un texto no restringido separado en tokens y se utiliza el algoritmo de Viterbi para hallar la secuencia de estados (etiquetas o categorías) más probable para la secuencia de palabras observada en el texto de entrada. Para las palabras que intervienen en la especialización se procede a cambiar la etiqueta de la especialización por la original.

2.4. Aplicación en textos reales

Para evaluar el desempeño de este etiquetador se utilizó el mismo corpus de textos en español extraídos de las páginas web de periódicos argentinos, corpus que está formado por 277 oraciones y 7911 palabras.

Esta herramienta es consultable por Internet, http://www.dsic.upv.es/%7Efpla/demo.html, se introduce el texto a analizar en un cuadro y devuelve el texto etiquetado con el siguiente formato:

Texto a analizar: *Palabra1 palabra2 ... palabraT*

Texto etiquetado: *Palabra1_etiqueta1 palabra2_etiqueta2 … palabraT_etiquetaT*

Entre los errores observados en el proceso de etiquetado se pueden distinguir algunos que se presentan en forma sistemática. Específicamente se observa que al aparecer un verbo “haber” o “ser” lo etiqueta únicamente como verbo auxiliar, incrementando la frecuencia de etiquetado incorrecto. Otro error observado es la asignación de etiqueta de nombre propio a títulos enteros en los que alguna palabra que es un nombre propio.

De la aplicación se evidencia en forma global 288 errores en etiquetado, esto es, 288 palabras que NO obtuvieron en el proceso la etiqueta correcta. En términos relativos se traduce en un 96.36% de precisión global, con una confianza del 95% el intervalo correspondiente es (95.95%; 96.77%).

3. SISTEMA BASADO EN SMORPH/MPS

3.1. Smorph y MPS

El software Smorph, analizador y generador morfosintáctico desarrollado en el GRIL por Salah Aït-Mokhtar [2], realiza en una sola etapa la tokenización y el análisis morfológico. Puesto que es una herramienta declarativa, la información utilizada por Smorph está separada de la maquinaria algorítmica, por lo tanto es posible adaptarlo a distintos usos [3].

El módulo post-smorph MPS [4], es un analizador que recibe en entrada una salida Smorph (en formato Prolog) y puede modificar las estructuras de datos recibidos en la entrada y ejecuta dos funciones principales: la Recomposición y la Correspondencia. Estas dos funciones serán útiles para resolver las ambigüedades que resulten del análisis de Smorph.

Smorph compila, minimiza y compacta la información lingüística de modo que quede disponible en un fichero binario. Los códigos fuente se dividen en cinco archivos:

- Códigos Ascii
- Rasgos
- Terminaciones
- Modelos
- Entradas

Para obtener un etiquetador mediante la aplicación de estos dos sistemas se utiliza, además de lo que se detallará más adelante, toda la información declarada para determinar los límites de oraciones presentada en Beltrán (2007) [5]. A continuación se hará referencia a la información contenida en los archivos Smorph y MPS con el objeto de lograr un etiquetado de un texto y luego se desarrollará la utilización de MPS.

3.2. Archivo ASCII

En este archivo se especifican los caracteres separadores, las equivalencias entre mayúsculas y minúsculas, etc. Sin embargo, en esta aplicación, además se utiliza la función "ALTERS" prevista para declarar información en este archivo.

Dentro de la información declarada se detallan los separadores, los espacios y se introducen las mayúsculas. Esto hace que cuando Smorph no reconozca una entrada con alguna mayúscula, ésta será reemplazada por minúscula y, luego del reemplazo, si se encuentra en las entradas será ahora reconocida.

Dentro de este archivo también es posible definir pares de caracteres que pueden intercambiarse (alternarse). En este punto es importante destacar que mediante los ALTERS del archivo ASCII es posible definir las iniciales de nombres y cifras con puntos decimales o separadores de miles. Por ejemplo, cuando se declara "`A N .`" y Smorph encuentra una ocurrencia no reconocida en las entradas, va a intercambiar A por N y va a buscar si la nueva ocurrencia es reconocida. De esta manera, si en las entradas defino `"N.N."` como iniciales de nombre, cuando Smorph encuentre `"J.C."` va a intercambiar la `J` por `N` y la `C` por `N` y luego va a tener `N.N.` lo que sí tiene declarado en las entradas como iniciales de nombre y por lo tanto será reconocido como tal.

De la misma manera se pueden definir las cifras con separadores de miles o decimales.

3.3. Archivo de Rasgos

El fichero de **rasgos** contiene las etiquetas organizadas jerárquicamente, por ejemplo, nombre, adjetivo, etc., como así también las etiquetas que indican el tipo de nombre o bien los rasgos de concordancia, género y número.

3.4. Archivo de Terminaciones

En el archivo de terminaciones se declaran todas las terminaciones que se hace referencia en los modelos de flexión.

3.5. Archivo de Modelos de flexión

En el fichero de modelos, se introduce la información correspondiente a los modelos de flexiones morfológicas. Un modelo de flexión agrupa todas las flexiones de una misma clase de palabras y se describe asociando a un conjunto de terminaciones el correspondiente conjunto de definiciones morfológicas. En este punto se mostrará parte de los modelos utilizados para nombres, adjetivos y verbos.

3.5.1. Modelos para nombres

Retomando el trabajo de Solana-Rodrigo (2005) [6], se definieron 48 modelos para nombres. A continuación se presentan dos modelos para comprender la declaración de los mismos en el sistema.

Modelo n1: el modelo n1 definido en el archivo modelos de Smorph agrupa a todos aquellos nombres masculino singular que al concatenarle una "s" se obtiene su forma masculino singular. Esto se declara de la siguiente manera

```
@n1           -0
    +@          nom/masc/sg
    +s          nom/masc/pl .
```

Ejemplo: *barrilete, queso*

Modelo n4: el modelo n4 definido en el archivo modelos de Smorph agrupa a todos aquellos nombres que al sustraerle el último caracter y concatenarle una letra "o" se obtiene su forma masculino singular, al concatenarle una letra "a" se obtiene su forma femenino singular, mientras que al concatenarle "os" y "as" se obtiene respectivamente las formas masculino plural y femenino plural. La especificación del mismo es

```
@n4          -1
    +o          nom/masc/sg
    +a          nom/fem/sg
      +os          nom/masc/pl
      +as          nom/fem/pl .
```

Ejemplo: *abuelo, abogado*

3.5.2. Modelos para adjetivos

De la misma manera que para los nombres, se presentarán dos de los trece modelos definidos para los adjetivos.

Modelo a3: en este modelo se reúnen los adjetivos que son ambiguos respecto al género y que la forma plural se obtiene concatenándole una "s". Se recuerda que el lema coincide con la forma singular a la cual no se concatena nada.

```
@a3           -0
    +@          adj/_/sg
    +s          adj/_/pl .
```

Ejemplo: *amable*

Modelo a4: el modelo a4 agrupa a todos aquellos adjetivos que al sustraerle el último caracter y concatenarle una letra "o" se obtiene su forma masculino singular, al concatenarle una letra "a" se obtiene su forma femenino singular, mientras que al concatenarle "os" y "as" se obtienen respectivamente las formas masculino plural y femenino plural.La especificación del mismo es

```
@a4          -1
        +o              adj/masc/sg
        +a              adj/fem/sg
        +os             adj/masc/pl
        +as             adj/fem/pl .
```

Ejemplo: *bueno, lindo*

3.5.3. Modelos para verbos

El español presenta una compleja morfología verbal en los paradigmas llamados regulares y en mayor medida en los irregulares. Los verbos irregulares son aquellos que manifiestan en su flexión cambios vocálicos, consonánticos y acentuales en su raíz o tienen terminaciones distintas a las del paradigma regular. En el trabajo de Solana-Bonino-Valenti (2005) [7] se obtuvo la modelización de las fuentes declarativas en una herramienta de análisis y lematización automáticos de verbos del español que puede ser utilizada sobre textos reales. Ellos trabajaron los verbos de la primera conjugación (verbos en "ar") y luego fue extendido a todos los verbos en un trabajo posterior en el cual también se incluye en consideración de la morfología del verbo en español la segunda persona singular y plural de la variedad rioplatense.

Los modelos para verbos utilizados en este artículo corresponden a los definidos en dicho trabajo. A continuación se describe brevemente el enfoque utilizado.

La conjugación regular está expresada mediante un conjunto finito de terminaciones, cada una de las cuales está asociada a un conjunto de valores. En todos los casos, la raíz es el lema menos los dos caracteres de la terminación del infinitivo. Esa raíz va a concatenarse con cada una de las terminaciones. Dado que los verbos pueden ser irregulares en la raíz o en las terminaciones se determina el subconjunto de terminaciones regulares que se concatenan con raíces no regulares y el subconjunto de terminaciones irregulares que se concatenan con raíces regulares o no.

De esta manera se consigue extraer las terminaciones que no siguen la conjugación regular y por lo tanto su complemento es el conjunto de terminaciones que sí siguen las conjugaciones regulares. Esta metodología permite separar en un verbo irregular todo aquello que es regular y aquello que es irregular. Con este criterio y buscando siempre tratar de reunir los verbos en el menor número posible de tipos se han definido 63 modelos para verbos.

A continuación se describen algunos de los modelos especificados. Se presentan tres modelos, uno para verbos regulares y dos para verbos irregulares.

- Modelo para verbos regulares terminados en "ar": este modelo, llamado modelo v1, reúne a todos los verbos regulares con terminación "ar". Para comprender la lectura de la declaración en Smorph se describen algunas de las líneas.

Este modelo establece que al extraerle los dos caracteres correspondientes a la terminación y concatenarle:

- ✓ ar → se obtiene el infinitivo del verbo
- ✓ ando → se obtiene el gerundio del verbo
- ✓ ado/ados/ada/adas → se obtiene el participio del verbo en cada una de sus flexiones
- ✓ o → se obtiene la primera persona del singular del tiempo presente del modo indicativo
- ✓ as → se obtiene la segunda persona del singular del tiempo presente del modo indicativo
- ✓ a → se obtiene la tercera persona del singular del tiempo presente del modo indicativo

- Modelos 4 y 5: estos dos modelos se requieren para reunir aquellos verbos irregulares en los cuales la irregularidad de la raíz es del tipo vocálica. En el modelo 4 se especifica la parte irregular del verbo mientras que en el modelo 5 quedará determinado su complemento, esto es, la parte regular de estos verbos. Ejemplos de estos verbos son: *acertar* y *sonar*. Las entradas asociadas al modelo 4 presentan el siguiente formato:

```
acertar aciert    @v4 .
acertar     @v5 .
```

es decir, una entrada para la parte irregular y una entrada para su complemento. La primera entrada está asociada al modelo 4 el cual determina las terminaciones a concatenar para obtener aquellas formas en las cuales la raíz presenta irregularidad.

Ejemplo: **aciert** + **a** =acierta (tercera persona singular del tiempo presente del modo indicativo)

La segunda entrada está asociada al modelo 5 que constituye las formas flexionadas que mantienen regularidad. El mismo establece que se le extraen los dos últimos caracteres correspondientes a la terminación y se le concatenan ciertas terminaciones para obtener las formas flexionadas coincidentes con las de los verbos regulares terminados en "ar".

Ejemplos:

acert + **ar** = acertar (infinitivo)

acert + **amos** = acertamos (primera persona plural del tiempo presente del modo indicativo)

acert + **aba** = acertaba (primera persona singular del tiempo pretérito imperfecto del modo indicativo)

3.5.4. Modelos para verbos infinitivos y clíticos

El infinitivo de los verbos se obtiene mediante la concatenación de ar, er o ir a la raíz (según especifican los modelos para verbos). Sin embargo, para obtener el análisis de los verbos infinitivos con un clítico concatenado al final, se declaran otra vez los infinitivos de los verbos asociados a un nuevo modelo.

Este modelo especifica que cuando se encuentre el infinitivo + un clítico concatenado a continuación sin espacios se tiene infinitivo+clítico, por ejemplo "**amarlo**".

La especificación de este modelo es

```
@inf1                -0
    +me         infcl
    +te         infcl
    +nos        infcl
    …
    +les        infcl .
```

Ejemplo: ***amarse***, ***comerlo***, ***correrlas***, ***temerle***

Otra consideración hay que tener para el caso de un infinitivo con dos clíticos concatenados al final sin espacios en blanco, por ejemplo "**comérmelo**". El aspecto a tener en cuenta en este caso es que el infinitivo se encuentra acentuado ya que al adicionarse los dos clíticos la palabra resultante es esdrújula. Por este motivo es que se declaran nuevamente los infinitivos en las entradas, pero ahora acentuados (ár, ér o ír), asociados a un tercer modelo.

Ejemplos: ***corrértelo***, ***partírselo***, ***lavártelo***, ***enfriármelo***.

Esquema de entradas:

```
#Ejemplos de entradas para analizar infinitivo + 2 clíticos

abalanzar abalanzár   @inf2 .
abalear   abaleár     @inf2 .
abalizar  abalizár    @inf2 .
....
```

El modelo se especifica de la siguiente manera:

```
@inf2                -0
    +mete       infcl
    +melo       infcl
    +melos             infcl
…
    +senos             infcl .
```

3.6. Archivo de Entradas

En el archivo de entradas, se ingresan los ítems léxicos acompañados por un indicador del modelo correspondiente. El indicador del modelo es el responsable de relacionar las entradas con el archivo de modelos, donde se especifica la información morfológica, género y número y con el archivo de terminaciones, esto es, las terminaciones que se requieren en cada ítem.

3.7. Reglas de MPS

Cuando un texto es analizado por Smorph, utilizando la información lingüística mencionada en los puntos anteriores, devuelve un texto en el cual cada palabra presenta todos los análisis posibles. Por ejemplo, la palabra "*para*" presentará el resultado del análisis que corresponde a una preposición y al mismo tiempo presentará el resultado de un análisis que corresponde a las formas flexionadas del verbo *parar*:

```
'para'.
[ 'para', 'EMS','prep'].
[ 'parar', 'EMS','v', 'MODOV','ind', 'PERS','3a', 'NUM','sg', 'TPO','pres',
'TR','r', 'TC','c1', 'TDIAL','estrpi'].
[ 'parar', 'EMS','v', 'MODOV','imper', 'PERS','2a', 'NUM','sg', 'TPO','pres',
'TR','r', 'TC','c1', 'TDIAL','est'].
[ 'parir', 'EMS','v', 'MODOV','subj', 'PERS','1a', 'NUM','sg', 'TPO','pres',
'TR','r', 'TC','c3', 'TDIAL','estrpi'].
[ 'parir', 'EMS','v', 'MODOV','subj', 'PERS','3a', 'NUM','sg', 'TPO','pres',
'TR','r', 'TC','c3', 'TDIAL','estrpi'].
[ 'parir', 'EMS','v', 'MODOV','imper', 'PERS','3a', 'NUM','sg', 'TPO','pres',
'TR','r', 'TC','c3', 'TDIAL','estrpi'].
```

Esta ambigüedad presentada, como muchas otras observadas, se resuelven mediante la utilización del módulo Post Smorph, MPS, en el cual es posible definir reglas de correspondencia y recomposición que permitirán de alguna forma tomar una decisión.

En primer lugar se utilizan las reglas de correspondencias definidas en Beltrán (2007) donde se trató la delimitación de las oraciones.

La metodología de trabajo para construir las restantes reglas en este módulo fue la siguiente:

1. Se analizó un texto en Smorph conteniendo 8.541 palabras.
2. Se llevó el archivo resultante del análisis a una planilla de cálculo Excel en la cual se retuvo las etiquetas morfosintácticas asociadas a cada palabra. Puesto que sólo se consideró el primer nivel de etiquetas para el ejemplo anterior se retuvo por ejemplo la etiqueta de preposición y de verbo (dos etiquetas).
3. Se construyeron dos nuevas variables, la etiqueta anterior y posterior a cada palabra.
4. Mediante la utilización de tablas dinámicas de Excel se pudo observar las ambigüedades más frecuentes y las ocurrencias más frecuentes entre pares de etiquetas.
5. De la observación en 4 se definieron reglas útiles para resolver ambigüedades. Para resolver aquellos casos en los cuales no existe una regla se optó por la etiqueta más frecuente en cada caso. Por ejemplo, si se encontró una ambigüedad *nombre/verbo* imposible de resolver con las reglas de recomposición planteadas, se definió una regla de correspondencia asignando una de las dos etiquetas según la probabilidad de ocurrencia en el texto oficiante de "muestra".
6. Estas reglas, junto con el resto del análisis de Smorph presentado se aplicó al texto de análisis, ya utilizado en capítulos anteriores, para comparar los resultados con un etiquetador estadístico basado en modelos de Markov.

A continuación se presentan, a modo de ejemplo, dos de ellas:

1) AUXILIAR+PARTICIPIO+SUBORDINANTE:

```
S1 [ L1, 'EMS','aux'] S2[ L2, 'EMS','v', 'MODOV', 'part'] S3[ L3,
'EMS','sub']  --> S1+S2+S3 [ L1+L2+L3, 'EMS','aux_part_sub'].
```

Esta regla resuelve casos como:

```
'han'.
[ 'haber', 'EMS','aux', 'MODOV','ind', 'PERS','2a', 'NUM','pl', 'TPO','pres',
'TR','hi', 'TDIAL','estrpi'].
[ 'haber', 'EMS','aux', 'MODOV','ind', 'PERS','3a', 'NUM','pl', 'TPO','pres',
'TR','hi', 'TDIAL','estrpi'].

'dicho'.
[ 'dicho', 'EMS','nom', 'GEN','masc', 'NUM','sg'].
[ 'dicho', 'EMS','v', 'MODOV','part', 'TR','hi', 'TC','c3'].

'que'.
[ 'que', 'EMS','rel'].
[ 'que', 'EMS','sub'].
```

Resultado:

```
'han dicho que'.
[ 'haber dicho que', 'EMS', 'aux_part_sub' ].
```

2) AUXILIAR+PARTICIPIO:

```
S1 [ L1, 'EMS','aux'] S2[ L2, 'EMS','v', 'MODOV', 'part']  --> S1+S2 [ L1+L2,
'EMS','aux_part'].
```

Esta regla resuelve casos como:

```
'ha'.
[ 'haber', 'EMS','aux', 'MODOV','ind', 'PERS','3a', 'NUM','sg', 'TPO','pres',
'TR','hi', 'TDIAL','estrpi'].

'hecho'.
[ 'hecho', 'EMS','nom', 'GEN','masc', 'NUM','sg'].
[ 'hecho', 'EMS','v', 'MODOV','part', 'TR','hi', 'TC','c2'].
```

Resultado:

```
'ha hecho'.
[ 'haber hecho', 'EMS', 'aux_part' ].
```

Es importante recalcar que el orden en que se declaren las reglas es pertinente. Por ejemplo, si la regla 2 se declara antes que la regla 1 no va a encontrar auxiliares y verbos participios, es decir, `[ L1, 'EMS','aux']` y `[ L2, 'EMS','v', 'MODOV', 'part']` porque la regla 2 (si está antes) ya los ha juntado y transformado en `[ L1+L2, 'EMS','aux_part']`.

3.8. Aplicación en textos reales

Para evaluar el desempeño de este etiquetador se utilizó el mismo corpus de textos en español extraídos de las páginas web de periódicos argentinos que se utilizó para la segmentación de textos en oraciones y para la evaluación del etiquetador estadístico basado en bigramas.

En la aplicación se obtuvieron 66 errores en etiquetado, esto es, 66 palabras que no obtuvieron durante el proceso la etiqueta correcta. En términos relativos se traduce en un 99.17% de precisión global, con una confianza del 95% el intervalo correspondiente para el parámetro correspondiente a la precisión del sistema es (98.97% ; 99.37%).

4. DISCUSIÓN

Evaluando los intervalos de confianza obtenidos para la precisión global en ambos etiquetadores (Tabla 4.1) podemos observar que el etiquetador basado en conocimiento lingüístico se encuentra “por encima” del hallado para el sistema basado en bigramas. Esto sugiere, al considerar el texto de análisis como una “muestra aleatoria” de la población de textos con similares características, el sistema basado en SMORPH/MPS presenta una precisión global superior.

Tabla 4.1: Intervalos de confianza del 95% para la precisión global.

Sistema	Precisión	Límite inferior del intervalo (95% de confianza)	Límite superior del intervalo (95% de confianza)
Bigramas	96.36	95.95	96.77
Smorph/Mps	99.17	98.97	99.37

REFERENCIAS

[1] Pla, Ferrán; Molina, Antonio; Prieto, Natividad. "Evaluación de un etiquetador morfosintáctico basado en bigramas especializados para el castellano". Universidad Politécnica de Valencia. 2001.

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

[3] Aït-Mokhtar, Salah; Rodrigo Mateos, José Lázaro 1995 Segmentación y análisis morfológico de textos en español utilizando el sistema SMORPH. SEPLN Revista nº 17 pags 29-41.

[4] 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.

[5] Beltrán, Celina. Comparación de Sistemas para la Detección de Límites de Oraciones. Revista INFOSUR. Nro. 1 Junio 2007.

[6] Solana, Zulema; Rodrigo, Andrea. "El sintagma nominal núcleo". Publicación de Facultad de Filosofía y Letras de la Universidad Nacional de Cuyo. JALIMI 2005.

[7] Solana, Zulema; Bonino, Rodolfo, Valenti, Viviana. "Modelización de las fuentes declarativas en una herramienta de análisis y conjugación automáticos de verbos del español". Publicación de Facultad de Filosofía y Letras de la Universidad Nacional de Cuyo. JALIMI 2005.