Procesamiento del Lenguaje Natural, Revista nº 42, marzo de 2009, pp. 17-24 recibido 28-11-08, aceptado 02-03-09

Sistema de recomendación para un uso inclusivo del lenguaje*

Inclusive Language Recommendation System

Maria Fuentes, Lluís Padró, Muntsa Padró, Jordi Turmo y Jordi T. Carrera
Grupo de Procesamiento del Lenguaje Natural
Departamento de Lenguajes y Sistemas Informáticos
Universitat Politècnica de Catalunya
c/Jordi Girona, 1-3
08034 Barcelona
mfuentes,padro,mpadro,turmo,jcarrera@lsi.upc.es

Resumen: Sistema que procesa un texto escrito en castellano detectando usos del lenguaje no inclusivos. Para cada sintagma nominal sospechoso el sistema propone una serie de alternativas. El sistema permite también la adquisición automática de ejemplos positivos a partir de documentos que hagan un uso inclusivo del lenguaje. Éstos ejemplos seran usados, junto a su contexto, en la presentación de sugerencias.
Palabras clave: Lenguaje inclusivo, aprendizaje basado en ejemplos

Abstract: System to detect exclusive language in spanish documents. For each noun phrase detected as exclusive, several alternative are suggested by the system. Moreover, the system allows the automatic adquisition of positive examples from inclusive documents to be presented within their context as alternatives.
Keywords: Inclusive language, example based learning

1. Introducción

Hacer uso de un lenguaje inclusivo consiste en la selección de vocabulario y partículas de la lengua que permitan minimizar o eliminar las palabras que implican o parecen implicar la exclusión de un sexo. Por ejemplo *el personal de vuelo* o *la tripulación de cabina* es lenguaje inclusivo, mientras que *azafata* es claramente exclusivo (o sexista). De todas formas, para determinar el grado de lenguaje inclusivo a ser utilizado, (Wilson, 1993) remarca la importancia de tener en cuenta el sentido común si no se quiere que por las buenas intenciones se acabe sacrificando la prosa.

Existen varios manuales y herramientas que asisten a la producción de documentos inclusivos. Una de las primeras iniciativas en el estado español fue impulsada por el instituto de la mujer en el marco del proyecto *nombra.en.red* (Alario et al., 1995). En este proyecto se construyó un software de libre distribución, cuya base de datos fue creada siguiendo las sugerencias de usos alternativos que, en los años ochenta y noventa partieron, entre otros, del Consejo de Europa (Consejo Europa, 1986), del Institut Valencià de la Dona (Departamento Dona, 1987), del Instituto de la Mujer, de UNESCO y de la Conferencia de Naciones Unidas sobre las Mujeres de Pekín (Naciones Unidas, 1996).

Otra herramienta que podemos encontrar en la red es *la lupa violeta* (Factoria de Empresas, 2002). Fue diseñada para ser instalada en el procesador de textos Word, identifica los términos que pueden tener una utilización sexista y propone diferentes sugerencias. En la misma linea, recientemente se está comercializando *Themis* (The Reuse company, 2008), que explora archivos y sitios web en busca de usos exclusivos de la lengua ofreciendo alternativas de forma similar a los correctores ortográficos integrados en editores de textos.

Este artículo presenta el recomendador de alternativas inclusivas desarrollado en la UPC (Universidad Politécnica de Cataluña) para el proyecto Web con Género de la Fundación CTIC (Centro Tecnológico de la Información y la Comunicación)[1]. El sistema utiliza técnicas de aprendizaje basado en ejemplos y adquisición automática de ejemplos.

La sección 2 muestra una visión global del sistema, la 3 analiza el funcionamiento del sistema actual, la 4 propone posibles mejoras y la sección 5 concluye el artículo.

* Los autores desean mostrar su agradecimiento a Eulàlia Lledó y a Marta de Blas por la cesión de textos inclusivos, así como a Edgar Gonzàlez por facilitarnos su software de clustering.

[1] http://www.t-incluye.org

ISSN 1135-5948

2. *Arquitectura general*

Esta sección describe los componentes básicos de la arquitectura general del sistema. La Figura 1 presenta la interacción entre las cuatro componentes, cuyas funciones son: extraer características de un Sintagma Nominal (SN), filtrar el SN en función de si utiliza un lenguaje inclusivo o exclusivo, buscar en la Base de Datos (BD) ejemplos similares a cada SN exclusivo y por último presentar las mejores sugerencias inclusivas.

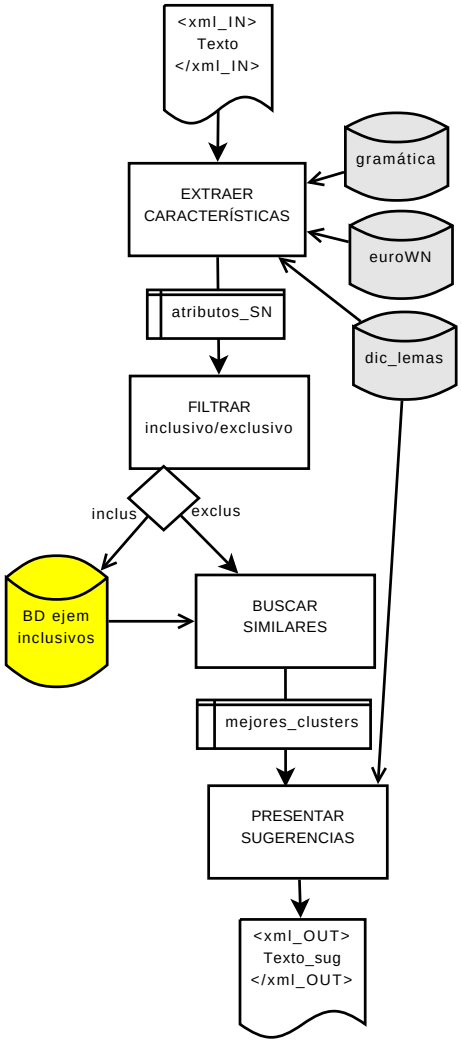

Figura 1: Componentes del recomendador.

Las dos funcionalidades básicas son:

- la detección de SNs susceptibles de hacer uso de lenguaje exclusivo y las correspondientes recomendaciones inclusivas.
- la adquisición automática de SNs inclusivos para la creación de forma automática de la BDs de ejemplos inclusivos.

El Cuadro 1 presenta un ejemplo de texto formateado como entrada del sistema, dividido en párrafos y codificado en XML. En el Cuadro 5 puede verse el formato de salida.

```
<DOC>
<INFO>
<URI>http::/www.un_dominio.es/una_pagina</URI>
<IP>192.168.2.243</IP>
<DATE>2998-03-13 11:34</DATE>
</INFO>
...
<P locator="136"type="texto">
3. La Junta Consultiva está constituida por el rector o la rectora,
que la preside; la secretaria general o el secretario general, que
lo es de la Junta, y cuarenta miembros más designados por el
Consejo de Gobierno, a propuesta del rector o la rectora, entre
profesoras o profesores e investigadoras o investigadores de
reconocido prestigio, de todos los ámbitos temáticos presentes
en la Universidad y de todos los que se considere oportuno,
acreditados por las correspondientes evaluaciones positivas de
acuerdo con la normativa vigente, ocho de los cuales, al menos,
deben ser externos a la Universidad Politécnica de Cataluña.
</P>
...
<P locator="164"type="texto">
4. A efectos de esta elección, la comunidad universitaria se
considera dividida en los cuatro sectores siguientes:
</P>
<P locator="165"type="texto">
a)Profesorado doctor de los cuerpos docentes universitarios.
b)Personal docente e investigador, excluido el correspondiente
al sector a.
c)Estudiantes.
d)Personal de administración y servicios.
</P>
...
</DOC>
```

Cuadro 1: Documento de entrada.

El primer paso consiste en extraer una serie de características (atributos) de cada SN.

En segundo lugar se tendrán en cuenta SNs inclusivos, cuando el objetivo sea la adquisición de ejemplos y SNs exclusivos cuando el objetivo sea la recomendación. En el primer caso se almacenarán en la BD los ejemplos filtrados y sólo en el segundo caso será necesario buscar ejemplos inclusivos similares existentes en la BD para finalmente presentar las sugerencias más adecuadas.

2.1. Extraer características

El objetivo de esta fase es obtener una serie de características morfosintácticas y semánticas necesarias en la siguiente fase para determinar si un sintagma es inclusivo *los hombres y las mujeres*, exclusivo *los hombres*, o irrelevante *los coches y las motos*.

La información extraída en esta fase también será utilizada en la búsqueda de ejemplos similares, tanto para indexar los ejemplos inclusivos en la BD cómo para seleccionar las mejores alternativas a un SN detectado como exclusivo.

La parte superior del Cuadro 2 presenta un ejemplo de SN inclusivo en su contexto, *los hombres y las mujeres*, y la inferior los atributos asociados. El número de atributos varía en función de las características del sintagma

nominal. Los atributos contienen información sobre lemas, formas, etiquetas morfológicas (en el ejemplo *parole*), información semántica (*sense*), etiquetas sintácticas (*label, multiple*).

En él se desarrollan algunos aspectos relacionados con la violencia: sus significados, los modos en que **hombres y mujeres** se posicionan ante la misma, las causas de la violencia ejercida específicamente contra las mujeres y el papel que juega la socialización de niñas y niños en la formación de conductas violentas.
atributos: lemma=y form=y parole=CC HasDoubleForm=false lemma1=hombre form1=hombres parole1=NCMP000 HasDoubleForm1=true senses1=0:07391044 0:05957670 0:07392506 0:01967203 0:07331418 0:07392045 1:06951621 1:00017954 1:00004123 1:01966690 1:07602853 1:06951621 2:00004123 2:00003731 2:00002086 2:01964914 2:07356184 2:00004123 3:00003731 3:00002086 3:00001740 3:00001740 3:01402712 3:00004123 3:00003731 3:00002086 4:00001740 4:00001740 4:01378363 4:00003731 4:00002086 4:00001740 4:00001740 5:00995974 5:00001740 5:00001740 6:00990770 7:00008019 8:00002086 9:00001740 lemma2=mujer form2=mujeres parole2=NCFP000 HasDoubleForm2=false senses2=0:07684780 1:06948278 2:00004123 3:00003731 3:00002086 4:00001740 label=sn-doble multiple=true

Cuadro 2: SN y las características extraídas

HasDoubleForm indica que un lema tiene forma para ambos géneros. Este atributo será también cierto en palabras masculinas que tienen contraparte femenina, pero que no comparten lema con ella y por tanto no son detectables vía diccionario, como *hombre*.

Esta fase utiliza la librería Freeling[2] (Atserias et al., 2006), que proporciona varios analizadores del lenguaje: análisis morfológico, etiquetado gramatical, análisis sintáctico superficial, detección y clasificación de entidades nominales y anotación semántica basada en WordNet (Vossen, 1998).

Un SN puede estar formado por varios nombres y cada uno de ellos puede a su vez tener varios sentidos. La información semántica asociada se ve reflejada en los atributos *sense*, Cuadro 2. La Figura 2 presenta parte de la información semántica asociada a los conceptos "hombre" y "mujer". Según WordNet mientras *mujer* tiene un único significado *hombre* puede tener varios y ambas palabras tienen por hiperónimo el concepto persona.

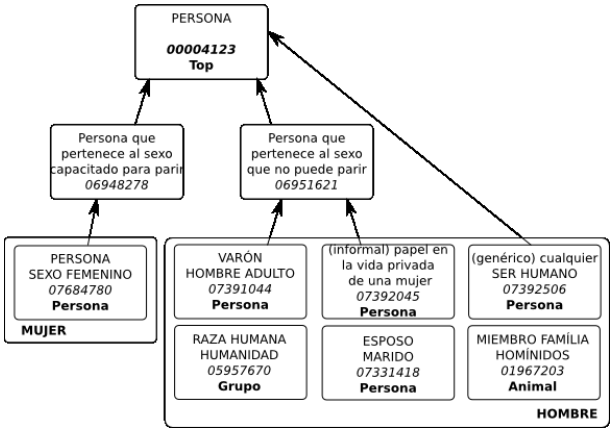

Figura 2: Representación semántica de hombre y mujer (sense1 y sense2 en Cuadro 2)

Referente al análisis sintáctico, para el recomendador se ha creado una gramática de SNs específica y se ha modificado el diccionario para que palabras cómo *príncipe* y *princesa* tengan el mismo lema.

2.2. Filtrar

El componente Filtrar puede considerarse como un clasificador de SNs. La Figura 3 presenta el árbol de decisión que se aplica para identificar si un SN es inclusivo (*CORRECTO*), exclusivo (*INCORRECTO*), irrelevante (*DESCARTAR*) o multiple (*DESMONTAR*).

En esta fase se aplican una serie de patrones que combinan información sintáctica con información semántica. Sintácticamente se tiene en cuenta si el SN es doble o sospechoso y semánticamente se tiene en cuenta si la palabra tiene una relación de hiperonimia con *persona* o *grupo social*.

La regla por defecto sería que si un SN hace referencia a una persona o grupo social en masculino que tiene contraparte femenina y esta no aparece reflejada se detecta cómo incorrecto, si aparece se detecta cómo correcto.

Para los casos a los que no se puede aplicar la regla por defecto o requieren un tratamiento especial para desvincularlo de la información que tiene o deja de tener WordNet se ha creado una serie de listas. A continuación se describe cada lista y el Cuadro 3 presenta las

[2]http://garraf.epsevg.upc.es/freeling/

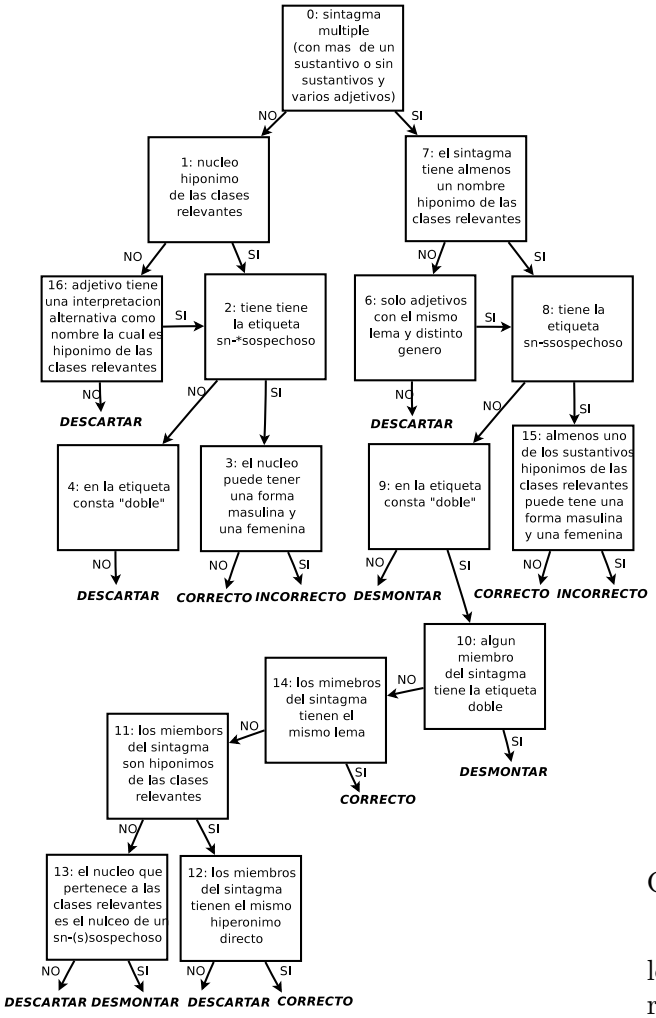

Figura 3: Representación del árbol de decisión para filtrar SN inclusivos o exclusivos.

palabras especiales
hombre varón macho padre papá papa padrino marido caballero patrono obispo cardenal poeta jinete judío primero segundo tercero último penúltimo amo capellán albañil
palabras no relevantes
par tipo sector curso seminario tribunal nombre corazón factor amor circo pueblo estado contacto región elemento compromiso animal negocio extremo conferencia servicio encuentro periódico ejército encuentro colegio consejo departamento instituto ejemplo cuerpo cabo centro congreso simposio espectáculo cielo reparto cuadro diario modelo banco capítulo campamento país conjunto éxito régimen bloque monstruo montón comedor imperio talento club partido palacio ministerio metro fantasma horario pájaro comité reino municipio ángel ayuntamiento vehículo cariño clan cerebro as cristianismo editorial sol base maricón terror satélite violín baile bajo testimonio bicho máquina academia laboratorio aula taller clínica campo doble papel general desastre demonio ex nazi rayo grande moro movimiento círculo miembro parte alfabeto
palabras inclusivas
persona
nombres vacíos
persona equipo señor colectivo sindicato ramo órgano población clase comunidad mundo coto profesión personal público gente grupo habitante asociación
palabras genéricas
profesorado alumnado ciudadanía estudiantado electorado clientela vecindario funcionariado voluntariado abogacía afición presidencia tropa vicepresidencia gerencia jefatura secretaría asesoría alcaldía coordinación redacción autoría magistratura judicatura delegación descendencia audiencia proletariado burguesía chiquillería humanidad juventud infancia adolescencia tesorería ingeniería ministerio consistorio tripulación pasaje consultoría auditoría notaría tutoría conserjería empresa directiva

Cuadro 3: Palabras con tratamiento especial.

palabras que contienen inicialmente.

La lista **palabras especiales** contiene lemas de palabras masculinas que tienen una palabra femenina, pero que no comparten lema con ella y por tanto no son detectables vía diccionario. Por ej. "niños" es una palabra masculina que comparte lema con "niñas", que es femenina. Palabras como "hombres" no tienen esta característica, dado que su correspondiente femenino ("mujer" en este caso) tiene un lema diferente.

El sistema usa información semántica extraída de WordNet para determinar si una palabra puede referirse a personas o a colectivos, que son conceptos clave para la identificación de ejemplos correctos o incorrectos en cuanto a género. Algunas palabras tienen sentidos poco frecuentes que caen en esas categorías (p.e. "un tipo" o "un par" pueden referirse a una persona, "un tipo majo", "un Par del Reino", "estar con sus pares (sus iguales)". Así mismo, "curso" o "sector" pueden referirse a un grupo o colectivo ("el curso de 5º son unos gamberros", "el sector del metal esta en huelga"). La lista **palabras no relevantes** contiene lemas de palabras para las que el sistema debe ignorar los sentidos persona/colectivo que puedan tener, ya que son poco habituales. Eso evita la inclusión en la BD de muchos ejemplos irrelevantes, corriendo el riesgo de descartar ejemplos relevantes en las pocas ocasiones en que esas palabras constituyan ejemplos a detectar.

La mayoría de palabras con género morfológico femenino o bien se refieren a objetos o a animales hembras (silla, casa, gata, gallina, ...) o a personas de sexo femenino (niña, amiga, ...). En el primer caso, no son relevantes para el tratamiento del lenguaje exclusivo. En el segundo, se considera que el deseo era referirse a una/s persona/s de sexo femenino y por tanto, no se detecta como sintagma incorrecto ni tampoco como sintagma candidato a sugerencia. Las palabras en la lista **palabras inclusivas** (como p.e. "persona") son excepciones a esta regla, y deben ser consideradas candidatas a sugerencia aunque sean morfológicamente femeninas.

La lista **nombres vacíos** contiene aquellos nombres que se refieren a una persona o colectivo, pero que en el caso de llevar un adjetivo, es ése el que aporta la información relevante (p.e. "persona usuaria" es relevante para "usuario", o "equipo directivo" lo es para "directivos" o "director").

La lista **palabras genéricas** contiene palabras que se refieren a colectivos, pero que en WordNet no aparecen como tal.

2.3. Buscar similares

El sistema utiliza una BD de ejemplos inclusivos indexada para que el acceso a los ejemplos sea eficiente utilizando técnicas de clustering o agrupación de ejemplos. Lo que significa que se agrupan los ejemplos según su parecido, para facilitar su posterior recuperación por similitud. En concreto se accede a los clusters o conjuntos de ejemplos con menor distancia (valor entre 1 y 0). La distancia entre ejemplos se calcula aplicando la siguiente fórmula:

$$d = 1 - ((Pla * Sla + Pf * Sf + Ple * Sle + Ps * Ss + Pp * Sp)/Pnormaliza)$$

donde Sla, Sf, Sle, Ss y Sp son respectivamente las similitudes entre las etiquetas sintácticas, las formas, los lemas, los sentidos y las etiquetas morfológicas y Pnormaliza es la suma de los pesos de cada similitud: Pla 0.1, Pf 3, Ple 5, Ps 8 y Pp 1.

Se ha utilizado una implementación de Clustering Jerárquico Aglomerativo (Jardine y Sibson, 1971). Como distancia inter-grupo hemos utilizado "Unweighted Pairwise Group Method using Arithmetic Averages" (Zhao y Karypis, 2002). Una vez el dendrograma está construido, el número óptimo de clusters se determina usando Silhouette (Rousseeuw, 1987). Se selecciona la profundidad del árbol cuyos clusters obtienen un mayor valor Silhouette.

Adicionalmente el sistema tiene dos parámetros relacionados con la construcción de los clusters:

Número mínimo de clusters de ejemplos que se crearan. El algoritmo decide automáticamente el número óptimo de grupos, pero en algunos casos el criterio de decisión puede no obtener un valor satisfactorio. En estos casos, se usa el número de clusters especificado en esta opción.

Número máximo de ejemplos en un cluster. Se usa en el proceso de decisión del número de clusters. Si el corte óptimo supone crear un cluster de tamaño mayor al valor dado en esta opción, se busca otro valor óptimo que no viole esta restricción.

Los ejemplos de la BD se agrupan en clusters y para cada cluster se elige un ejemplo representante (*medoide*). El Cuadro 4 presenta los ejemplos que forman el cluster representado por el medoide *una educadora o un educador*.

637: del equipo educativo
917: una educadora o un educador
1065: la persona así educada
1771: educadoras y educadores
1798: como persona educadora
1803: educadoras o educadores
1804: de un equipo educativo
4292: la persona educadora
4698: educadoras/es
medoide: 917
num.ejemplos: 9

Cuadro 4: Ejemplo de cluster y su *medoide*.

Para evitar comparar cada vez la distancia del SN tratado a todos los ejemplos de la BD únicamente los medoides son tenidos en cuenta en la selección del conjunto de clusters que se encuentran a menor distancia. En esta fase, se calcula la distancia del SN tratado con el medoide de cada cluster en la BD.

2.4. Presentar sugerencias

La selección de las sugerencias para un ejemplo incorrecto requiere el paso previo de selección de los clusters más prometedores. En esta última fase sólo se analizan las posibles sugerencias que contienen los mejores clusters, evitando así un recorrido exhaustivo de toda la BD. De entre las sugerencias analizadas, se seleccionan las más parecidas al ejemplo incorrecto, siempre que se encuentren dentro de un margen de similitud, y procurando que sean lo más variadas posible.

A continuación se describen los parámetros que controlan la búsqueda y selección de sugerencias:

Número máximo de sugerencias que dará el recomendador. Puede dar menos si no hay bastantes candidatos lo suficientemente cercanos al ejemplo incorrecto.

Umbral de distancia a partir del cual no se consideran las sugerencias, aunque no se haya

alcanzado el *número máximo de sugerencias*. La distancia equivale a 1-similitud, por lo que un umbral 0.55 implica que no se propondrán sugerencias con una similitud inferior a 0.45. Una distancia demasiado baja excluye ejemplos interesantes pero semánticamente alejados (p.e. ciudadanos de ciudadanía)

Número de clusters más cercanos al SN incorrecto a explorar para la selección de sugerencias. Si el valor es muy alto, se pierde eficiencia ya que se explora gran parte de la BD. Este valor controla el porcentaje de la BD que se explora en cada consulta. Si la BD tiene muchos clusters, que este valor sea alto, no necesariamente significa una gran perdida de eficiencia, y en cambio, garantiza que se encuentren los mejores ejemplos. Un valor de 1 puede funcionar bien con una BD rica en ejemplos. Un valor de 2 o 3 introduce cierta flexibilidad en la búsqueda que puede mejorar los resultados en ejemplos que quedan a medio camino entre dos grupos.

Umbral de igualdad. Para aumentar la variedad de las sugerencias, el recomendador omite los candidatos si son muy parecidos a alguno ya propuesto. (ej: si en la lista ya figura "los profesores y las profesoras", se omitirá "los profesores o las profesoras"). Este umbral es la similitud mínima que deben tener dos ejemplos para ser considerados "demasiado parecidos". Cuanto mayor es el valor, más estricta es la comparación (más parecidos se permite que sean los ejemplos de la lista final). Si el valor baja, menos estricta es la comparación (se consideraran parecidos ejemplos con mayores diferencias).

En caso que no se seleccione ningún ejemplo candidato, siempre que sea posible, se genera automáticamente una sugerencia sin contexto a partir del diccionario, "alcaldesa y alcalde" para el SN que contiene "alcalde".

3. Análisis del funcionamiento

Para mostrar lo que se puede esperar del sistema, analizaremos las sugerencias ofrecidas a una serie de SNs detectados como *exclusivos*, ver Cuadro 5.

Para permitir el acceso a un mayor número de ejemplos en la decisión de si dos SNs son similares no se tiene en cuenta las preposiciones, ni la mayoría de veces los adjetivos. El sistema propone usos inclusivos parecidos asociados a un contexto. Por ejemplo el primer SN detectado como incorrecto, "los usuarios", sólo podría ser remplazado directamente por "toda persona usuaria", sin embargo sin tener en cuenta la preposición y adecuando el número, se puede considerar que todas las sugerencias aportan información útil.

A veces información relevante de la sugerencia queda en lo que sería la zona de contexto: "los colectivos *de homosexuales*" o "los/*las trabajadores/as*". En el primer caso, sintácticamente se trata de dos sintagmas: un sintagma nominal "los colectivos" y uno preposicional "de homosexuales". Si el sistema considerara que es un solo sintagma, entonces se filtrarían como correctos ejemplos como "las personas del bar de la esquina" o "[avisar a] las personas del peligro que corren". Cuando aparece '/' el etiquetado sintáctico no acaba de ser del todo correcto, puesto que su uso es gramaticalmente discutible.

La calidad de los documentos de los que se han extraído los ejemplos es básica. Por ejemplo en el caso de "para médicos, enfermeras, dietistas y otros profesionales" aparecen usos exclusivos de lenguaje, sin embargo "para médicos" aparece como ejemplo porque en el diccionario utilizado "médico" es una palabra que se puede referir a ambos géneros. Así pues, la primera sugerencia propone la eliminación del artículo para que sean incluidos profesionales de ambos sexos. No obstante lo que sucede es que el sistema no comprueba que el contexto sea inclusivo. Por esta razón "otros profesionales", exclusivo, aparece en el contexto de un ejemplo positivo.

La tercera alternativa que se da a "los médicos" es "doctor o doctora", sugerencia aceptable, aunque el significado en el contexto dado no sea sinónimo de "médico". Cada aparición de un SN es almacenado una sola vez en la BD, independientemente de su contexto o de si pueda tener varios significados.

El sistema no suele presentar sugerencias a los adjetivos. En el tercer párrafo, "los turistas alemanes" se da alternativas para "los turistas" proponiéndose quitar el determinante para incluir tanto turistas femeninos como masculinos. Será necesario la posterior supervisión de las concordancias en el texto final.

Si el contenido de la BD ha sido creado a partir de la adquisición automática de ejemplos es recomendable una supervisión de su contenido. Ya que puede ocurrir que los contextos sean poco significativos o como en el caso de "de mujer o por los investigadores", sugerencia propuesta a "los investigadores", se haya almacenado como inclusivo un ejem-

<P locator="1"type="texto"> **Los usuarios** del recinto se manifiestan **en contra de los homosexuales**. </P>

<P locator="1"type="texto">
<SN end="14"fac="1.0"start="2"id="1">**Los usuarios**</SN>
<L_SUG id="1"> <SUG sim="1.0"id="1">
... situación, que desorienta <*EJ*>*a los colectivos usuarios*</*EJ*> de los servicios formativos, se simplifica ...
</SUG> <SUG sim="0.9490392648287383"id="2">
... seguridad de redes y sistemas o Informática <*EJ*>*de usuario/a*</*EJ*> o Programador/a de aplicaciones ...
</SUG> <SUG sim="0.9172932330827067"id="3">
... guardar el rastreo de lo que hace <*EJ*>*toda persona usuaria*</*EJ*> de Internet durante un mínimo de ...
</SUG> </L_SUG>
<SN end="71"fac="1.0"start="42"id="2"> **en contra de los homosexuales**</SN>
<L_SUG id="2"> <SUG sim="0.9941520467836257"id="1">
Represión franquista y lucha de los colectivos <*EJ*>*de homosexuales*</*EJ*> y transexuales Fernando Olmeda, ...
</SUG> <SUG sim="0.9422156790577841"id="2">
... especialmente en las escuelas, como estos pares: <*EJ*>*homosexual*</*EJ*>/heterosexual; femenino/masculino; ...
</SUG> <SUG sim="0.9364319890635678"id="3">
... homosexual es no ser ya ni mujer ni hombre, como si <*EJ*>*la persona homosexual*</*EJ*>renunciara a su ...
</SUG> </L_SUG>
Los usuarios del recinto se manifiestan en contra de los homosexuales.
</P>

<P locator="2"type="texto"> **Los trabajadores** optan por unirse **a los médicos**. </P>

<P locator="2"type="texto"> <SN end="17"fac="1.0"start="1"id="3">**Los trabajadores**</SN>
<L_SUG id="3"> <SUG sim="0.9941520467836257"id="1">
... Comisiones Obreras cuando pide la equiparación de los/<*EJ*>*las trabajadores/as*</*EJ*> del sexo al resto ...
</SUG> <SUG sim="0.993815730994152"id="2">
Tanto <*EJ*>*los trabajadores y trabajadoras*</*EJ*> propuestos por la Fundación Universidad de Oviedo, como ...
</SUG> <SUG sim="0.9294976571864444"id="3">
... fácil podría ser cambiar el mundo si <*EJ*>*toda persona trabajadora*</*EJ*> donara una unidad monetaria ...
</SUG> </L_SUG'>
<SN end="48"fac="1.0"start="35"id="4"> **a los médicos**</SN>
<L_SUG id="4"> <SUG sim="0.9941520467836257"id="1">
... necesita para conducir un servicio de entrenamiento <*EJ*>*para médicos*</*EJ*>, enfermeras, dietistas y otros ...
</SUG> <SUG sim="0.9406850459482038"id="2">
... enfermedad todavía muy desconocida incluso <*EJ*>*para el colectivo médico*</*EJ*>, y es objeto de miles de ...
</SUG> <SUG sim="0.48454469507101083"id="3">
... personal docente e investigador con el grado <*EJ*>*de doctor o doctora*</*EJ*>, que ha de constituir, ...
</SUG> </L_SUG>
Los trabajadores optan por unirse a los médicos.
</P>

<P locator="3"type="texto"> **Los turistas alemanes** serán premiados **por los investigadores**. </P>

<P locator="3"type="texto">
<SN end="24"fac="1.0"start="3"id="5"> **Los turistas alemanes**</SN>
<L_SUG id="5"> <SUG sim="0.9941520467836257"id="1">
Llegan a nuestro país <*EJ*>*como turistas*</*EJ*> y la consiguiente exigencia de visados al poco de su estancia ...
</SUG> </L_SUG>
<SN end="63"fac="1.0"start="41"id="6"> **por los investigadores**</SN>
<L_SUG id="6"> <SUG sim="0.9941520467836257"id="1">
Bienestar reune en Madrid (CSIC, Serrano 117) <*EJ*>*investigadores/as*</*EJ*> de más de 30 paises.
</SUG> <SUG sim="0.9472398946083156"id="2">
... bagaje colectivo <*EJ*>*como grupo investigador*</*EJ*>está formado por el pensamiento crítico de teóricos ...
</SUG> <SUG sim="0.7571929824561403"id="3">
temas <*EJ*>*de mujer o por los investigadores*</*EJ*> que (¿casualmente?) son mayoritariamente mujeres, ...
</SUG> </L_SUG>
Los turistas alemanes serán premiados por los investigadores.
</P>

<P locator="4"type="texto"> **El director** se reune con **el alcalde**. </P>

<P locator="4"type="texto"> <SN end="13"fac="1.0"start="2"id="7">**El director**</SN>
<L_SUG id="7"> <SUG sim="0.9941520467836257"id="1">
Cada vez es más habitual ver 'informático/a' o '<*EJ*>*director/a*</*EJ*>', pero seguimos encontrándonos con ...
</SUG> <SUG sim="0.9410175981620718"id="2">
Las decanas y los decanos y <*EJ*>*las directoras y los directores*</*EJ*> de las unidades deben elaborar y someter ...
</SUG> <SUG sim="0.8624011007911937"id="3">
Una profesora me contó que el año pasado <*EJ*>*el equipo directivo*</*EJ*> de su instituto decidió gastarse todo ...
</SUG> </L_SUG>
<SN end="37"fac="1.0"start="23"id="8"> **con el alcalde**</SN>
<L_SUG id="8"> <SUG sim=1.0"id="1">
<*EJ*>*alcaldesa y alcalde*</*EJ*>
</SUG> </L_SUG>
El director se reune con el alcalde.
</P>

Cuadro 5: Ejemplo de párrafos no inclusivos y las sugerencias ofrecidas por el recomendador.

plo que en realidad es exclusivo.

Por último, toda sugerencia podrá ser susceptible de error, ya que el sistema no tiene manera alguna de saber si el texto se está refiriendo a un varón concreto, por ejemplo, si el alcalde es un hombre no tiene sentido sugerir "alcaldesa y alcalde".

4. Trabajo futuro

La definición final del contenido de las listas de palabras utilizadas para la configuración definitiva del sistema, así como la ampliación de los ejemplos positivos de la BD, se está llevando a cabo en la Fundación CTIC.

Una mejora del sistema consiste en tener en cuenta todas las partículas del SN, ya que por el momento básicamente se tienen en cuenta nombres. Los adjetivos sólo se tienen en cuenta si el núcleo del SN es un nombre que aparece en la lista *nombres vacíos*.

Tratar los pronombres nos permitiría detectar ejemplos como "estamos todos y to-

das" o malos usos como "contacte con nosotros". De todas formas, por el momento el sistema tampoco trata verbos, por lo que ninguna construcción con clíticos, "contactarnos", puede ser detectada como correcta.

Retocar la gramática y el extractor de características mejoraría el tratamiento de SNs especialmente complejos cómo "de nuestras hijas e hijos, amigas y amigos y colegas". El extractor actual sólo obtiene información de tres elementos por SN: palabra "," o "conjunción", palabra1 y palabra2.

Refinar el árbol de decisión con que se implementa el filtrado de SNs en el tratamiento de SNs dobles permitiría que no se filtrasen cómo ejemplos positivos SNs del estilo "de mujer o por los investigadores".

El estudio de la calidad del contexto en la extracción de SNs inclusivos también significaría una mejora, evitando incluir ejemplos, como el anteriormente mencionado, cuyo contexto contiene "otros profesionales".

Referente a la presentación de sugerencias, no se tiene en cuenta si el SN viene precedido por una preposición o no. Se podría estudiar la posibilidad de poner las preposiciones en la zona de contexto. De manera que las sugerencias a "Los usuarios" serían: "los colectivos usuarios, usuario/a y toda persona usuaria", en lugar de "a los colectivos usuarios, de usuario/a y toda persona usuaria".

Queda como trabajo futuro la detección y corrección de SNs que excluyan a personas de sexo masculino, "azafatas" o "enfermeras".

5. Conclusiones

El sistema presentado tiene dos funcionalidades básicas: la recomendación de un uso del lenguaje inclusivo y la adquisición automática de ejemplos inclusivos a partir de textos considerados correctos.

El sistema utiliza aprendizaje basado en ejemplos. Por ello, la calidad de las recomendaciones es fuertemente dependiente de la calidad y cantidad de ejemplos previamente almacenados en la Base de Datos, aunque cómo toda aplicación de inteligencia artificial tiene asociado un cierto grado de error. Por esta razón el recomendador debe ser considerado como un asistente a la escritura de textos inclusivos y no como un corrector de textos exclusivos.

Bibliografía

Alario, Carmen, Mercedes Bengoechea, Eulalia Lledó, y Ana Vargas. 1995. En femenino y en masculino. Madrid: Ministerio de Trabajo y Asuntos Sociales.

Atserias, Jordi, Bernardino Casas, Elisabet Comelles, Meritxell González, Lluís Padró, y Muntsa Padró. 2006. Freeling 1.3: Syntactic and semantic services in an open-source nlp library. En *Proceedings of the fifth international conference on Language Resources and Evaluation (LREC 2006), ELRA*, Genoa, Italy.

Consejo Europa. 1986. Igualdad de sexos en el lenguaje. Comisión de terminología en el Comité para la igualdad entre mujeres y hombres del Consejo de Europa.

Departamento Dona. 1987. Recomendaciones para un uso no sexista de la lengua. Consellería de Cultura, Educación y Ciencia de la Generalitat Valenciana.

Factoria de Empresas. 2002. La lupa violeta. http://www.factoriaempresas.org/productos yresultados/lupavioleta/lanzador.swf.

Jardine, N. y R. Sibson. 1971. *Mathematical Taxonomy*. John Wiley and Sons, Inc.

Naciones Unidas. 1996. Declaración de pekín y plataforma para la acción. IV Conferencia mundial sobre las mujeres, Pekín.

Rousseeuw, Peter. 1987. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20:53–65, November.

The Reuse company. 2008. Themis. http://www.themis.es.

Vossen, Piek. 1998. Eurowordnet: A multilingual database with lexical semantic networks. Dordrecht. Kluwer Academic Publishers.

Wilson, Kenneth G. 1993. *The Columbia Guide to Standard American English*. Columbia University Press.

Zhao, Y. y G. Karypis. 2002. Evaluation of hierarchical clustering algorithms for document datasets. En *Proceedings of the Eleventh International Conference on Information and Knowledge Management (CIKM'02)*, páginas 515–524.