Revista Española de Documentación Científica
39(2), abril-junio 2016, e133
ISSN-L:0210-0614. doi: http://dx.doi.org/10.3989/redc.2016.2.1236

NOTAS Y EXPERIENCIAS / *NOTES AND EXPERIENCES*

Titulación automática de preguntas en encuestas electorales

Carolina Gallardo*, Jesús Cardeñosa**

*Departmento de Sistemas de Información,
Escuela Técnica Superior de Sistemas Informáticos. Universidad Politécnica de Madrid
**Grupo de Validación y Aplicaciones Industriales. ETS de Ingenieros Informáticos. Universidad Politécnica de Madrid
Correos-e: cgallardo@eui.upm.es, carde@fi.upm.es

Recibido: 11-11-2014; 2ª versión: 16-03-2015; Aceptado: 28-07-2015.

Cómo citar este artículo/Citation: Gallardo, C.; Cardeñosa, J. (2016). Titulación automática de preguntas en encuestas electorales. *Revista Española de Documentación Científica*, 39(2): e133. doi: http://dx.doi.org/10.3989/redc.2016.2.1236

Resumen: Este artículo describe el trabajo realizado para la generación automática de los títulos de las preguntas pertenecientes a las encuestas de opinión que existen en las bases de datos del CIS (Centro de Investigaciones Sociológicas). Dentro del contexto del CIS, el título de una pregunta debe cumplir dos requisitos: desde el punto de vista de la forma, debe ser gramaticalmente correcto y tener un estilo similar a los ya existentes; y, desde el punto de vista del contenido, debe albergar el tema de la pregunta y las distintas categorías de respuesta. Estas restricciones en cuanto a la forma y al contenido de los títulos desaconsejan el uso de técnicas empleadas en problemas similares, como el resumen automático o aprendizaje automático con corpus de entrenamiento, a favor de una metodología basada en el análisis y conocimiento del dominio. Para ilustrar el análisis y la estrategia de resolución del problema seguidos, hemos seleccionado las preguntas relacionadas con temas electorales, debido a la importancia estratégica y a la especialización del CIS en este tipo de encuestas. Se describe en detalle el procedimiento seguido y la evaluación de los resultados, valorando tanto los aspectos cualitativos como los cuantitativos. La evaluación muestra que el 88,73% de los títulos generados cumplen estrictamente con los requisitos de forma y contenido impuestos por el CIS, lo que supone un ahorro en el trabajo manual del personal cualificado de la institución.

Palabras clave: Minería de textos; recuperación de información; filtrado; clasificación; resumen y visualización; titulación automática; extracción de información; encuestas de opinión.

Automatic Titling of Election Survey Questions

Abstract: This paper describes the work carried out for automatically generating titles for questions included in the opinion polls contained in CIS databases (Centro de Investigaciones Sociológicas – Spanish Center of Sociological Research). In the context of CIS, the title of a question should meet two requirements: from the point of view of form, it has to be grammatically correct and similar in style to existing ones; from the point of view of content, it must contain the subject of the question and the different options for answering. These conditions for form and content of titles discourage the use of techniques used in similar problems, such as automatic abstracting or machine learning with a training corpus, but rather favor a methodology based on an analysis and knowledge of the domain. To illustrate the analysis and the resolution strategy of the problem, we have selected a set of questions related to elections, due to their strategic importance and to CIS's own specialization in opinion polls. The process followed and the subsequent evaluation of results are discussed in detail, with an assessment of both qualitative and quantitative aspects. The evaluation shows that 88.73% of the generated titles are in strict accordance with CIS's requisites on form and content, resulting in reduced time spent by the institution's qualified personnel on manual work.

Keywords: Automatic titling; information extraction; opinion polls; abstracting.

1. INTRODUCCIÓN

La amplia variedad de elementos sobre los que trabajan los sistemas de búsqueda, como noticias de prensa con encabezados y títulos a modo de campos de búsqueda activos, preguntas para la creación de encuestas de opinión o materiales audiovisuales, representan diferentes contextos en los que el *título* es un concepto clave que sirve para guiar y efectuar la búsqueda. Es decir, un título puede usarse para clasificar, efectuar búsquedas o indizar recursos. Por ello, no es raro que exista interés tanto en identificar y extraer títulos de documentos sin formato como en intentar generarlos automáticamente en una amplia variedad de unidades textuales: desde colecciones de documentos, emisiones de eventos deportivos, documentos individuales o noticias (generación de titulares).

El CIS (Centro de Investigaciones Sociológicas) es una institución oficial que lleva encuestando a la sociedad española desde 1960. Aparte de liderar la investigación sociológica en España, una de sus tareas es distribuir la información relativa a los resultados de las encuestas a periodistas, investigadores o a cualquier ciudadano. Por lo tanto, requiere de un sistema que busque, indexe y recupere unidades de información como son las preguntas provenientes de miles de encuestas de finales de los años sesenta hasta la fecha. Este artículo describe el procedimiento seguido para automatizar el proceso de generación de títulos en todas las preguntas sin título que aparecen en las bases de datos de encuestas del CIS. La titulación automatizada formaba parte de una serie de tareas que perseguía la homogeneización y unificación de todas las bases de datos de esta institución, cuyo principal objetivo era generar títulos de alta calidad a través del uso de tecnologías rápidas y baratas.

En principio, si consideramos un título como un resumen condensado, el proceso de titulación puede considerarse asimismo como una tarea de resumen, de manera que las técnicas desarrolladas en esta área se podrían aplicar al problema del titulación. De hecho, en sus inicios, la titulación automática se ha abordado mediante técnicas pensadas para el resumen de textos, como las denominadas "extractivas" (Goldstein y otros, 1999). No obstante, existen diferencias notables entre un título y un resumen que hacen que se descarte la idea de aplicar técnicas extractivas a la tarea de generación de títulos. En primer lugar, los títulos están sujetos a fuertes restricciones de forma, estilo y extensión que cada ámbito de aplicación impone. En (García-Gutiérrez, 2014), se apunta la dificultad del tratamiento de los titulares de noticias debido a la presencia de elementos retóricos en los mismos, mientras que las características que definen a un buen resumen no dependen tanto del ámbito de aplicación. Asimismo, es importante resaltar la relevancia cognitiva del título frente al resumen: se trata de una pieza clave en un espacio de información (Martínez-Ávila y otros, 2014) con funciones de clasificación, indexación o incluso como elemento orientado a capturar atención del lector.

Teniendo en cuenta cómo el ámbito de aplicación determina la forma y la idoneidad de un título, cualquier técnica (como la extractiva) que no incorpore las restricciones particulares de un ámbito en particular y que pueda producir resultados incoherentes e incorrectos resulta inadecuada para la tarea de generación de títulos. En este sentido, la titulación ha desarrollado sus propias técnicas y metodologías, evolucionando hacia modelos de aprendizaje automático (Jin y Hauptmann, 2001), dependientes del ámbito de aplicación y de la existencia de corpus de entrenamiento anteriores de documentos que están titulados.

Los sistemas basados en el corpus de entrenamiento y en el aprendizaje automático son más robustos por norma general que los basados en modelos extractivos y producen títulos de mejor calidad (Jin y Hauptmann, 2002), ya que tienen en cuenta el estilo y las restricciones de extensión impuestas en cada ámbito. El uso de técnicas de aprendizaje automático explota el conocimiento subyacente del corpus formado por documentos titulados y disponibles con el objetivo de emplear ese conocimiento para adaptar los nuevos títulos generados a los ya existentes.

En el área de resumen, pero, especialmente, en la titulación, se ha hecho énfasis en los enfoques estadísticos y probabilísticos, en contraposición con los modelos simbólicos basados en la lingüística (Spärck Jones, 2007). El trabajo aquí descrito comparte con los que están basados en las técnicas de aprendizaje automático la dependencia del ámbito, por lo que busca generar títulos *parecidos a los ya existentes*, pero depende más de los basados en el conocimiento del dominio que en los empíricos. En esencia, seguimos un enfoque basado en patrones, un enfoque clásico en Inteligencia Artificial que se ha aplicado en una amplia variedad de tareas, como la extracción de eventos en la web (Hung y otros, 2010) , la población de ontologías (Liu y otros, 2011), la respuesta de preguntas y la extracción de información (Cui y otros, 2007; Spasic y otros, 2010), por mencionar algunas de las aproximaciones más recientes.

Hemos considerado la titulación de preguntas como una tarea de extracción de la información más que como un problema de resumen de contenidos. Mientras que en (Gallardo y otros, 2011)

damos una imagen de conjunto del problema del ámbito, este artículo detalla y se centra en las preguntas electorales, como motivo de su importancia cualitativa en el CIS.

2. PREGUNTAS ELECTORALES: ANÁLISIS Y TIPOLOGÍA

Una **pregunta** es un texto corto cuyo objetivo es recopilar el valor de una o más variables sociológicas y sus distintas opciones de respuesta. Una variable sociológica puede variar desde información objetiva del entrevistado (como su situación laboral, nivel de educación o edad) a la opinión del mismo acerca de un tema o persona pública en concreto. Por lo tanto, el título de una pregunta debe contener dos tipos de información:

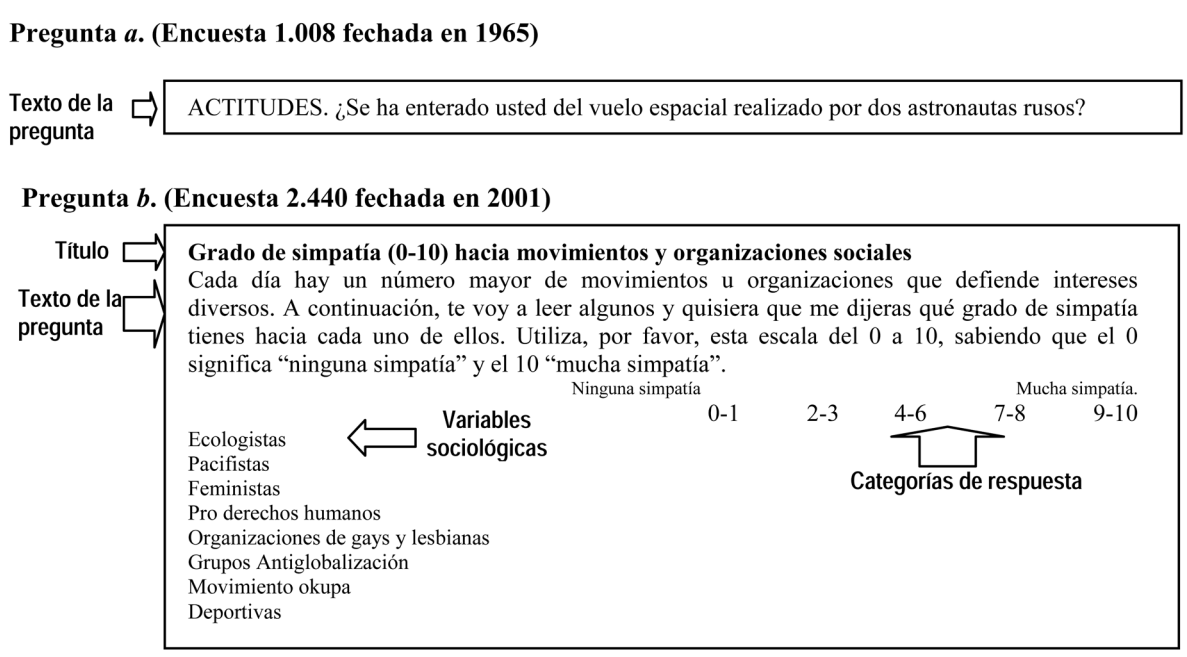
- Palabra clave: generalmente, una palabra o un término que sugiera las categorías de respuesta de la pregunta. Por ejemplo, "Grado", "Evaluación", "Aprobación", "Preferencia", etc.
- Tema: una expresión nominal que resuma los contenidos de la pregunta ("simpatía hacia movimientos sociales", "labor del presidente de la Comunidad Autónoma", "diferentes alternativas de organización del estado".

La Figura 1 muestra dos ejemplos de preguntas. La Pregunta *a* es un ejemplo de estructura sin título, perteneciente a un estudio de 1965, y la Pregunta *b*, bastante más reciente, está dotada de una estructura mucho más rica.

Debido a la coexistencia de formatos distintos, el CIS decidió homogeneizar sus bases de datos y procesar automáticamente las preguntas cuyas partes no están identificadas para obtener entidades estructuradas, incluyendo la tarea de titular las preguntas. A la hora del procesamiento, había un total de 84.769 preguntas almacenadas en la base de datos, de las que 39.257 habían sido tituladas manualmente por el personal del CIS y usadas como modelo de referencia para definir distintos tipos de títulos y 45.512 preguntas no habían sido tituladas, así que se convirtieron en nuestro objetivo.

La hipótesis en la que trabajaremos es la de intentar automatizar el procedimiento que realiza el personal del CIS, teniendo en cuenta las restricciones de su propio ámbito, e intentar aplicar estas restricciones en las preguntas sin título. De esta forma, aseguramos tanto homogeneidad como uniformidad en los nuevos títulos en comparación con los ya existentes. Para hacerlo, nos vamos a centrar en el análisis de la relación de naturaleza predominantemente lingüística entre una pregunta y su título. Este trabajo sigue por tanto un enfoque basado en el conocimiento del dominio.

Figura 1. Dos preguntas modelo con niveles de estructuración diferentes



Para realizar un análisis de dominio, la primera tarea que se abordó fue la descripción y clasificación de las preguntas y sus títulos, prestando atención a aspectos como el número total de títulos distintivos y sus frecuencias, los temas más frecuentes o la estructura gramatical de los mismos. Estos aspectos se describen a continuación.

2.1. Temas presentes en los títulos

Los temas relacionados con las preguntas electorales resultan ser uno de los más significativos y demandados por la población y los medios especializados. Los temas que definen a las preguntas electorales son los siguientes:

- Afinidad y simpatía con respecto a los partidos políticos
- Conocimiento sobre las afiliación política de los líderes políticos
- Conocimiento de los eslóganes electorales
- Escala de la ubicación ideológica de partidos políticos
- Fidelidad en el voto en las elecciones
- Intención de voto (*se incluyen varias configuraciones y distintos tipos de votantes)*
- Momento de voto o decisión de abstención
- Razones de votar X /de no votar *(se incluyen varias configuraciones)*
- Participación electoral común
- Participación en unas elecciones dadas *(en el pasado o futuro)*
- Partido por el que no votarán
- Partido que prefieren que gane
- Predicción electoral
- Prioridades (partido o candidato) en la decisión del voto
- Recuerdo de voto *(distinguiendo los distintos tipos de votantes)*
- Las opiniones del encuestado acerca del partido votado en las elecciones pasadas

2.2. Estructuras gramaticales de los títulos

Aunque las preguntas electorales son bastante homogéneas y uniformes en su formulación, se requiere ver cómo se relacionan las unas con las otras desde un punto de vista lingüístico, considerando aspectos como qué contenidos están presentes en el título y cómo se expresan (paráfrasis, interpretación, resumen creativo, reformulación completa o repetición exacta). Se observan tres tipos de relaciones lingüísticas:

A. El título es una interpretación de las variables o de las categorías de respuesta

Es frecuente que en algunas preguntas, como las multivariables (con distintas posibilidades de respuesta para una misma pregunta), el título sea un resumen o incluso una interpretación de todas las variables que se incluyen en la pregunta. Veamos el ejemplo siguiente:

Título: Opinión sobre el voto diferenciado a partidos dependiendo del tipo de convocatoria

Me gustaría que me dijera, ahora, si Ud. está más bien de acuerdo o más bien en desacuerdo con cada una de las siguientes frases. --

De acuerdo En desacuerdo

- En las elecciones autonómicas es mejor votar a un partido propio de la Comunidad Autónoma
- Lo lógico es votar siempre al partido que está más cerca de las propias ideas, independientemente de que las elecciones sean generales o autonómicas.
- En cualquier tipo de elecciones lo más importante son los candidatos.
- En las elecciones autonómicas es mejor votar a un partido distinto al que esté en el Gobierno Central, para evitar que gobierne el mismo en todas las instituciones.

En este caso, el título es un resumen de las distintas opciones de voto que expresa cada variable (votar al mismo partido, a partidos locales, etc.). Puesto que la tarea de titulación se abordará como un proceso de extracción de la información, no idearemos técnicas que requieran la comprensión o interpretación del texto para obtener el tema de la pregunta.

B. El título es una nominalización o paráfrasis

En este tipo de relación, parte del título es una sustantivación o paráfrasis de un fragmento de la pregunta con ligeras variaciones. En la siguiente pregunta, la expresión inicial "Visión" es una sustantivación y simplificación del sintagma verbal "ver en televisión":

Título: Visión en televisión de propaganda electoral

P21 ¿Ha visto Ud. por televisión algún espacio de propaganda electoral de algún partido o coalición?—

- Sí (Pasar a P21a)
- No
- N.C.

 Rev. Esp. Doc. Cient., 39(2), abril-junio 2016, e133. ISSN-L: 0210-0614. doi: http://dx.doi.org/10.3989/redc.2016.2.1236

C. El título es un fragmento exacto de la pregunta

Por último, el tema del título puede ser un fragmento exacto del texto de la pregunta, tal como ocurre en la siguiente pregunta:

Título: Valoración de los resultados electorales en las últimas elecciones autonómicas y municipales

P.23 En conjunto, y con independencia de sus preferencias personales, ¿cómo valora los resultados de las últimas elecciones municipales y autonómicas?–

- Más bien positivamente
- Más bien negativamente
- N.S./ N.C.

Intuitivamente, el primer tipo de relación (el tema como una interpretación de variables) requerirá un tipo de procesamiento distinto de B (nominalización) y C (fragmento). La interpretación de variables implica, en principio, un procesamiento de semántica profunda, mientras que, para el resto de relaciones, el problema de titulación se podría solucionar mediante un proceso de extracción de la información que sea relevante para el título.

3. PROCESAMIENTO Y RESULTADOS

La metodología y estrategia que se va a seguir consiste en un número de pasos basados en los hallazgos de la fase previa de análisis, a saber:

1. **Preprocesamiento de la pregunta**. En esta fase, los archivos de entrada se transforman en un formato adecuado para su procesamiento.
2. **Definición de reglas para el procesamiento de preguntas y la composición del título**. El desarrollo de esta tarea está determinado por las observaciones y conclusiones derivadas de la fase de análisis.
3. **Resultados y evaluación**. Cuando se generen los títulos, estos deben ser evaluados. En este paso, el personal del CIS tiene un papel importante.

3.1. Preprocesamiento de la pregunta

En esta tarea, se limpian los caracteres no imprimibles y se identifican y delimitan claramente las preguntas, con el objetivo de facilitar el procesamiento de cadenas. Para este problema específico, todas las preguntas sin *título* fueron extraídas y llevadas a un archivo de texto con la siguiente información:

- Un identificador numérico.
- El texto completo de la pregunta, que puede contener otros elementos, como instrucciones para el encuestador, múltiples variables y categorías de respuesta.

3.2. Reglas para el procesamiento de preguntas electorales

En esta tarea se codifican las reglas para la extracción de elementos y la producción de títulos. Las preguntas electorales pueden requerir un tipo de información concreta, como el partido votado o el reconocimiento de eslóganes, u opiniones y valoraciones relativas a las elecciones. Cada tema especificado en la sección 3 define un tipo de pregunta de modo que se definen una o varias reglas con un objetivo doble: identificar el tipo de pregunta y extraer la información requerida para la composición del título. Además, hay otro elemento: el *tipo de elecciones* (regionales, municipales, nacionales o europeas), que ha de ser identificado en el texto de la pregunta e incorporado en el título. Por tanto, la estrategia general para estas preguntas es:

a. Identificación del tipo de pregunta.
b. Identificación de los elementos de la pregunta que sugieran la palabra clave y el tema.
c. Identificación del *tipo de elecciones*.
d. Composición del título con los elementos encontrados.

A continuación, se describen en detalle las reglas para uno de los tipos más paradigmáticos, uniformes y frecuentes de preguntas sobre los tipos de elecciones: *Predicción electoral,* junto con la regla para el *Tipo de elecciones*.

Desde un punto de vista sociológico y haciendo frente a elecciones futuras, resulta más interesante encuestar sobre la intención de voto o predecir el partido ganador. El siguiente conjunto de reglas sirven para titular las preguntas que tienen que ver con la predicción de los partidos ganadores en elecciones futuras. La predicción tiene dos configuraciones principales: la predicción del partido ganador y la del candidato ganador. Similarmente a las reglas anteriores, si el tipo de elecciones aparece en la pregunta, esta información se debe incorporar en el título. Estas reglas se configuran de la siguiente manera:

```
IF pregunta =~ /(cree|piensa) (Vd|Ud|usted)
que <nombrePartidoPolitico> (tiene|tendría)
muchas <seqWords> ganar/)

THEN {
```

SI ∋ <eleccion> → titulo = "Previsión electoral para <nombrePartidoPolitico> de ganar $eleccion";

ELSE titulo = "Previsión electoral para <nombrePartidoPolitico>"

}

Esta regla establece que:

SI el texto de la pregunta contiene el verbo "cree" o "piensa", seguido opcionalmente del pronombre "Usted" o cualquiera de sus variantes, a su vez, seguido del partido político, junto con el verbo "tiene" o "tendría", "muchas" más una secuencia indefinida de palabras y, finalmente, del verbo "ganar".

Entonces, el título propuesto es "Previsión electoral para <nombrePartidoPolítico>".

Donde nombrePartidoPolitico se reconoce por medio de expresiones regulares de complejidad variable. La pregunta siguiente se corresponde con esta regla:

TÍTULO: Previsión electoral para el PSC/PSOE de ganar elecciones autonómicas

PREGUNTA: ¿Cree usted que el PSC/PSOE tiene muchas posibilidades, bastantes posibilidades, o muy pocas posibilidades de ganar las próximas elecciones?

También es posible que se mencione el nombre de un político en lugar de un partido, como ocurre en la siguiente regla:

SI pregunta =~ /(cree|piensa) (Vd|Ud|usted) que <NombrePropio> , <seqWords> , (tiene|tendría) <seqWords> posibilidad <seqWords> ganar/

SI ∋<eleccion> → titulo = "Previsión electoral para <NombrePropio> de ganar $eleccion";

ELSE titulo = "Previsión electoral para <NombrePropio>";

La siguiente pregunta se corresponde con esta regla:

TÍTULO: Previsión electoral para Juan Hormaechea de ganar elecciones autonómicas

PREGUNTA: P11 ¿Cree Vd. que Juan Hormaechea, encabezando la lista de su nuevo partido, tiene muchas posibilidades, bastantes, pocas o ninguna posibilidad de ganar las próximas elecciones autonómicas?

Ambas reglas dan cuenta de las distintas configuraciones del tema *Predicción de las elecciones.* El tipo de elecciones se reconoce de manera análoga, mediante la siguiente expresión regular:

IF pregunta =~ /(elecciones) $_{$2}$((\s\s?[a-zÁÉÍÓÚÑáéíóúñ]{2,})+)/

elección = "elecciones$2"

Por último, es posible que el título generado presente errores, como los puntos suspensivos y los pronombres personales *Usted/tú*, que no deberían aparecer en los títulos. Una vez que se procesen las preguntas y se generen los títulos, se aplican las reglas para la corrección de errores.

Hemos explicado en detalle las reglas para la titulación de un tema paradigmático dentro de las preguntas electorales. El resto de temas se trata de forma muy similar. Cada tema, por norma general, tiene una o dos reglas, con un total de 32 reglas responsables de la totalidad de los títulos de las elecciones generados.

El flujo de trabajo de procesamiento aquí descrito se muestra gráficamente en la Figura 2.

3.3. Resultados y evaluación

La estimación de la calidad de los resultados de este trabajo se ha hecho considerando tanto la calidad como la cobertura de los títulos generados.

3.3.1. Evaluación de la calidad de los títulos generados

En la valoración de los aspectos cualitativos de los títulos, el personal del CIS tuvo un papel importante. Basándose en su propia especialidad, clasificaron los títulos generados en tres tipos:

- **Títulos correctos:** aquellos que no presentan errores gramaticales y que contienen la palabra clave y el tema.
- **Títulos imprecisos:** aquellos que, a pesar de ser gramaticalmente correctos, no representan adecuadamente el tema de la pregunta. Por ejemplo, el siguiente título es gramaticalmente correcto pero ambiguo/impreciso en su significado:

  Previsión electoral para este partido de ganar las elecciones
- **Títulos incorrectos:** aquellos que presentan errores gramaticales, parecen inconclusos o son ilegibles.

Teniendo en cuenta esta clasificación, procedimos a evaluar los títulos generados de preguntas electorales, que alcanzaban un total de hasta 3.458 títulos. Para ello, se escogió una muestra aleatoria (con un margen de error del 5% y un nivel de confianza del 95%) de 346 preguntas. La Tabla I muestra los resultados de la muestra, así como su inferencia para el total de la población, basada en esos resultados.

Fig. 2. Esquema de la metodología

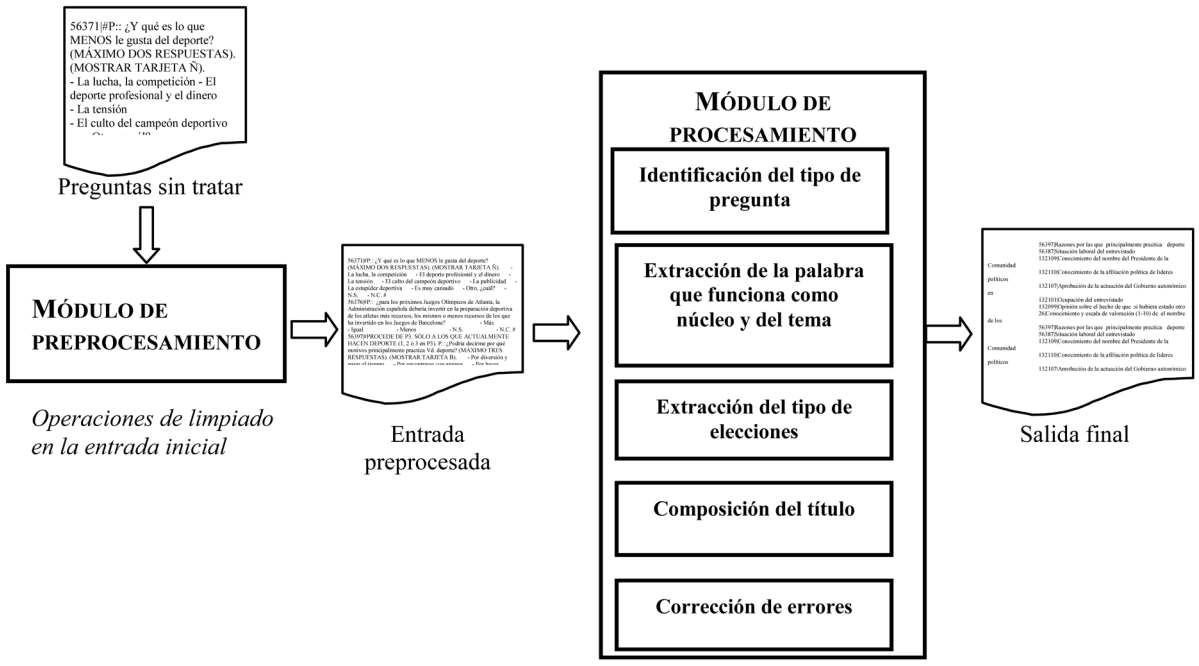


Tabla I. Evaluación de los títulos de las elecciones generados

Tipo	Resultados de la muestra	% Presencia de cada tipo en la muestra	Ocurrencias estimadas en la población
Correcto	307	**88,73**	3.068
Incorrecto	3	**0,87**	30
Impreciso	36	**10,40**	360
TOTAL	**346**	**100%**	**3.458**

Así pues, estimamos que las 32 reglas que cubren la totalidad de preguntas electorales generan correctamente el 88,73% de los títulos de las elecciones. Se observa asimismo que el número de títulos incorrectos es bastante bajo.

3.3.2. Evaluación de la cobertura de los títulos generados

La evaluación de la cobertura tiene como objetivo estimar la proporción de preguntas electorales para las que se ha generado un título (3.458) en relación al número total de preguntas electorales existentes en el corpus de preguntas no tituladas. Sin embargo, desconocemos a priori el número total de preguntas electorales existentes en el conjunto total de preguntas (45.512).

Para estimar el número de preguntas electorales dentro del total de preguntas (45.512), escogimos una muestra aleatoria de 381 preguntas, con un nivel de confianza del 95% y un margen de error del 5%. Los resultados se muestran en la Tabla II.

Considerando los resultados de la estimación, la cobertura se definiría de la siguiente manera:

$$\frac{\#títulos\ electorales\ generados}{\#títulos\ electorales\ existentes} = \frac{3.458}{5.495} = 62{,}93\%$$

Los resultados de la evaluación se resumen en la tabla III.

La cobertura no es muy alta, pero aquello en lo que nos hemos centrado en el trabajo no es en la obtención de un sistema de elevada cobertura, sino en ver si el esfuerzo llevado a cabo en el diseño, la codificación y la aplicación de las reglas era rentable en cuanto a costes y tiempo en relación con el método de revisión manual que el personal del CIS ha estado empleando hasta este momento.

Tabla II. Estimación del número total de preguntas electorales en la población entera

Tipo	Resultados de la muestra	% Presencia de cada tipo en la muestra	Ocurrencias estimadas en el total de población
Electoral	46	12,07%	5.495
No electoral	335	87,93%	40.017
TOTAL	381	100,00%	45.512

Tabla III. Tabla resumen de resultados

Títulos correctos	Títulos incorrectos	Preguntas sin título
88,73 % *(3069 títulos)*	11,27% *(389 títulos)*	37,07% *(2037 preguntas)*
62,93% *(3458 títulos)*		

El ahorro efectivo en costes y tiempo se ha estimado como altamente considerable, así que este enfoque se ha adoptado en conjunción con la revisión manual para la tarea global de la titulación de preguntas.

No obstante, merece la pena observar las razones por las que no se ha conseguido una mayor cobertura. El primer hecho notable consiste en que las preguntas que no se cubren se corresponden con las preguntas más antiguas de la base de datos, cuando la normalización de preguntas era un tanto baja, o incluso inexistente. En concreto, las situaciones observadas que inciden directamente en la falta de cobertura y de corrección fueron:

a) Aparición de nuevos tipos no identificados en la fase de análisis.

b) Insuficiencia de reglas. En este caso, las reglas codificadas no consiguieron capturar la información relevante.

c) Preguntas cuyo título es una interpretación o resumen de sus categorías de respuesta. Este tipo de preguntas quedaban fuera del ámbito del trabajo.

La presencia de nuevos tipos (caso a) sugiere que el corpus inicial de preguntas tituladas y sin titular no se comporta de forma parecida. Las preguntas sin titular son más antiguas que las tituladas, las cuales fueron tomadas como referencia para el análisis del ámbito. Si tenemos en cuenta que las encuestas del CIS reflejan el comportamiento de, en este caso, la sociedad española, puede que estemos encontrándonos con un signo claro de cómo la sociedad española se ha estado familiarizando con el sistema electoral actual, de tal modo que las preguntas electorales a principios de la democracia son más irregulares y "dependientes del momento y del contexto" (preguntaban acerca de situaciones novedosas que no se repetían en encuestas anteriores, solo en las posteriores). En contraposición, en los 90, tanto el sistema electoral como las encuestas electorales llevadas a cabo por esta institución están consolidas.

El caso b, sin embargo, revela que hay ciertas insuficiencias en las reglas diseñadas. El momento en que se realizó este trabajo, no se consideraba necesaria la expansión del número de reglas, principalmente debido a que un pequeño incremento en precisión podría significar un esfuerzo en mayores proporciones. Por otro lado, la cobertura y calidad de los títulos producidos con las reglas existentes fueron consideradas de gran utilidad.

4. CONCLUSIONES

Este trabajo describe el proceso que llevó a la titulación automática de las preguntas electorales del CIS. Este tipo de preguntas es uno de los más buscados y pedidos y su gestión y reutilización en las encuestas es esencial debido al gran número de situaciones electorales que pueden surgir.

Por otra parte, hemos seguido un enfoque directo, bastante lejos de sistemas complejos de difícil trato por parte de los responsables de estas tareas en sus instituciones. La solución diseñada tenía requisitos iniciales de eficacia, simpleza y, muy notablemente, de eficiencia, esto es, en un tiempo limitado con recursos limitados.

La evaluación de los resultados resultó ser bastante satisfactoria, ya que el 88,73% de los títulos generados eran correctos, con el consecuente ahorro en trabajo manual. Creemos que este enfoque, si no puede ser explotado directamente, establece las guías para la tarea de la titulación de preguntas de un modo rápido y directo.

5. REFERENCIAS

Cui, H.; Kan M.; Chua T. (2007). Soft pattern matching models for definitional question answering. *ACM Transactions on Information Systems, vol. 25*(2), pp. 1-30. http://dx.doi.org/10.1145/1229179.1229182

Gallardo Pérez, C.; Cardeñosa, J. (2011). Knowledge extraction for question titling. In *Proceedings of the 9th international conference on Flexible Query Answering Systems* (FQAS'11), Springer-Verlag, Berlin, Heidelberg, vol. 7022, pp. 119-127. http://dx.doi.org/10.1007/978-3-642-24764-4_11

García Gutiérrez, A. (2014). Análisis documental de noticias de prensa en sistemas de información factual. *Revista Española de Documentación Científica*, vol 37(2). http://dx.doi.org/10.3989/redc.2014.2.1094

Goldstein, J.; Kantrowitz, M.; Mittal, V.; Carbonell, J. (1999). Summarizing text documents: Sentence selection and evaluation metrics. *Proceedings of the 22Nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval,* Berkeley, California, USA. 121-128. http://dx.doi.org/10.1145/312624.312665

Hung, S.; Lin, C.; Hong, J. (2010). Web mining for event-based commonsense knowledge using lexico-syntactic pattern matching and semantic role labeling. *Expert Systems with Applications,* vol. 37(1), pp. 341-347. http://dx.doi.org/10.1016/j.eswa.2009.05.060

Jin, R.; Hauptmann, E. G. (2001). Headline generation using a training corpus. *Proceedings of the Second International Conference on Computational Linguistics and Intelligent Text Processing, CICLING*. Lecture Notes on Computer Science, Berlin: Springer-Verlag, vol. 2004: 208-215.

Jin, R.; Hauptmann, A. G. (2002). A new probabilistic model for title generation. *Proceedings of the 19th International Conference on Computational Linguistics, vol. 1.* http://dx.doi.org/10.3115/1072228.1072365

Liu, K.; Chapman, W. W.; Savova, G.; Chute, C. G.; Sioutos, N.; Crowley, R. S. (2011). Effectiveness of lexico-syntactic pattern matching for ontology enrichment with clinical documents. *Methods of Information in Medicine,* vol. 50(5), pp. 397-407. http://dx.doi.org/10.3414/ME10-01-0020

Martínez-Ávila, D.; San Segundo, R.; Zurian, F. (2014). Retos y oportunidades en organización del conocimiento en la intersección con las tecnologías de la información. *Revista Española de Documentación Científica*, vol. 37(3). http://dx.doi.org/10.3989/redc.2014.3.1112

Spärck Jones, K. (2007). Automatic summarising: The state of the art. *Information Process.Management, vol. 43*(6), pp. 1449-1481. http://dx.doi.org/10.1016/j.ipm.2007.03.009

Spasic, I.; Sarafraz, F.; Keane, J. A.; Nenadic, G. (2010). Medication information extraction with linguistic pattern matching and semantic rules. *Journal of the American Medical Informatics Association,* vol. 17(5), pp. 532-535. http://dx.doi.org/10.1136/jamia.2010.003657