Knowledge Extraction for Question Titling

Carolina Gallardo Pérez and Jesús Cardeñosa

Validation & Business Applications Research Group
Universidad Politécnica de Madrid–Spain
{carolina,carde}@opera.dia.fi.upm.es

Abstract. This article describes the work performed over the database of questions belonging to the different opinion polls carried during the last 50 years in Spain. Approximately half of the questions are provided with a title while the other half remain untitled. The work and implemented techniques in order to automatically generate the titles for untitled questions are described. This process is performed over very short texts and generated titles are subject to strong stylistic conventions and should be fully grammatical pieces of Spanish.

Keywords: titling, KE, summarization, opinion polls, subjective clustering.

1 Introduction[1]

The wide variety of elements that search systems can look for, like press news with headings and titles as active fields for search, questions for the creation of opinion polls, audiovisual materials (where their metadata can be longer than a mere title but shorter than a summary) or the media, represent different contexts where the title is a key concept to guide and perform the search. These contexts have used tools and techniques to generate titles and summaries with different results.

Although titling can be considered as a summarization task, there are some distinctive features between them, like the usually shorter length of titles as opposed to summaries, stronger stylistic conventions imposed over titles, and the different nature of the input texts to be summarized or titled.

There are two main approaches for the production of a summary, headline or title: to compose it from extracts of the input document (extractive summarization) or to compose it from an abstract of the document, identifying its central subject matter. Extractive summarization tries to identify the relevant sentences of a document. Thus summarization is viewed as a classification problem of relevant/non relevant sentences, which mainly relies on statistical knowledge [1]. To distinguish relevant from irrelevant sentences, several criteria can be used: position of sentences (where sentences heading or closing the paragraph are considered to be more relevant) is used in [2]; the presence of *signature words* (that can be defined by means of frequency measures like the tf-idf schema) is determinant in [3]; or sentence length combined

[1] The work reported in this article has been sponsored by CIS (Center for Sociological Research) of the Ministry of the Presidency of Spain.

with positional criteria and the presence of certain words is reported in [4]. Besides, since extractive techniques may produce incoherent or ungrammatical outputs, it can be the case that the generated summary or title is not required to constitute a completely grammatical expression [5, 6]. By far, these techniques are the most frequent approach for generating titles, headlines or summaries.

On the other hand, non-extractive trends in single-document summarization rely on knowledge-based techniques and tend to be domain-dependent approaches. For instance, a paradigmatic system like SUMMONS [7] is restricted to the summarization of news about terrorism. SUMMONS firstly extracts relevant information (like places, victims, authors, date, etc.) from texts using predefined templates. Then the extracted information is passed through a language generator module, which is also template-based. Other knowledge-based approaches make use of linguistic processing, like [8], together with domain knowledge [9], [10]. In any case, during the last decade there has been much less research and work in knowledge-based summarization (see [11] for a comprehensive review of the summarization task).

Due to their generality, it is difficult to find appropriate ad-hoc solutions based on these techniques to very specific problems, like the one presented in this paper: titling of a huge corpus of questions of the opinion polls carried out by the *Center for Sociological Research* (hereafter CIS) of Spain. CIS is a public institution with a long and stable tradition with its origins in the early 60ies. This means more than 50 years collecting sociological data from the Spanish society stored in different formats, means, databases and information supports. This institution decided to homogenize the structure of all their questions and surveys, a decision that involved the task, among others, of question titling, since titles would be used to identify similar questions in order to reuse them when composing new surveys based on previous ones or to establish temporal series of similar questions. Although all updating processes were initially manually performed by the CIS staff, beginning from the newest survey to oldest one, it turned out quite unmanageable and highly unproductive to manually review thousands of questions.

This article describes the methodology, specifically designed for this institution, to automate the process of question titling. The new generated titles should serve for the aforementioned search purposes. To do that, Information Extraction techniques have been applied in order to extract and identify the relevant and distinctive parts of the questions in order to build up the whole title. Although our proposal is based on domain-dependent criteria, it could be applicable to similar problems. The article is structured as follows: section 2 contextualizes this work and the problem to be solved; section 3 describes the preliminary analysis of the domain that ends up with a typology of titles; the resolution strategy is exemplified with a case study, developed in section 4; final remarks are stated in section 5.

2 Context

CIS carries out regular opinion polls to extract sociological data from the Spanish society. A sociological variable can vary from a specific piece of information about the interviewee (like labour situation, education level, social class, number of cars that the interviewee has, etc.) to the interviewee's opinion about a given issue, institution or public person. These opinion polls follow a fixed structure in order to elicit sociological variables and usually consist of a set of ordered questions.

A question can be viewed as a more or less complex structure with the following parts:

- A title or expression of the concept that underlies the whole question.
- The question text, that is, the exact wording of what is asked to the interviewee (includes an introduction, the question itself and instructions to the interviewer).
- Sociological variables (from one to many) that are covered by the question. They can coincide with the title or the title can be a grouping of the variables.
- The answer categories that establish the range and scope of the permissible answers to the question.

Our specific problem is defined by the necessity to assign a title to the questions that belong to surveys –opinion polls– of CIS. Besides, a title should meet two conditions:

a) It should contain the topic concept of the question (what the question is about)
b) It should imply the typology of answers categories of the question (yes/no question, multiple choice, scales and degree of evaluation).

When CIS begins with the updating processes, the situation is defined by the following figures:

- Number of questions in the database: 87221
- Number of manually titled questions: 39257
- Number of untitled questions: 47964 (from which, 1627 questions are dismissed since they resulted from an updating process)
- Erroneous questions: 150

As already said, the process of question titling is manually performed by experts. It is a creative process, like summarizing or translation, which heavily depends on the style and understanding of each person. Besides, for a specific question several alternatives can be posed and be acceptable. Thus, when facing the task of the automatic creation of title, we are proposing an automatic solution for a creative process.

3 Preliminary Analysis

The nature of the problem is determined by two main factors: a) titling is a creative process; and b) there already exists a big corpus of titled questions so that new titles have to be similar to the existent ones. Taking into account these determining factors, we proceeded to study the plausibility of the task.

The corpus of titled questions was thoroughly analyzed with a clear objective in mind: look for regularities in titles and their associated questions. The analysis focused on the linguistic features of titles from both a syntactic and a pragmatic point of view, highlighting aspects like the type of linguistic construction and the subjacent intention of the title. It also attempted to unveil determining aspects like the relation of the title with regard to the question.

Under such a formalist perspective, any thematic analysis fell out of the scope of this work (although CIS is specialized in pre- and post-electoral surveys and political issues and so it was expected to find many regularities and frequencies in the questions about these topics).

3.1 Setting the Work

This section depicts the analysis of existent titles and its associated questions. The main aspects that are analyzed are the type of linguistic construction (noun phrases, quotations, existence of paraphrases) and the degree of subjectivity/objectivity of the implicit linguistic enunciation in titles.

Most titles are noun phrases headed by a noun followed by a prepositional phrase of diverse complexity. There is almost no variety in this syntactic configuration. What is really striking is the variety in the subjective or objective nature of the linguistic expressions: evaluation, classification, dichotomies establishment, election from within a list of options, or simply assertion of an objective piece of information. The study of the intrinsic features of titles delivers two broad categories of titles: *subjective titles* and *objective titles*.

3.1.1 Subjective Titles

Titles under this category express an interviewee's judgement of any sort about a given topic. The judgement can be an approval, rejection, preference, evaluation, etc. of a topic or person. The type of judgement is explicitly expressed in the title, together with the object of the judgement.

In general, the structure of subjective titles follows the general schema of:

```
Type of judgement + Nexus + Topic
```

Where *type of judgement* is a word like "opinion", "preference", etc., *nexus* is the preposition or conjunction required by the head noun, and *topic* is the nominal group, clause or even quotation denoting what the question is about. Let's look at two particular examples:

TITLE: Opinion on the degree of interest of the central government in issues of the Valencian Community.
QUESTION: Q.24 Do you believe that the Central Government …?
- Tries hard to solve problems in the Valencian Community.
- Is interested in the economic progress of Valencia.
- Is fair in the sharing of the economical assets in Valencia.
-- A lot -- Sufficient – A little – Nothing

TITLE: ETA Terrorists' image
QUESTION: Which of these two statements do you most agree with?
-- ETA terrorists are criminals, heartless delinquents
-- ETA terrorists are idealist freedom fighters.

In these two examples, the titles turn out to be almost a personal interpretation of the question as well as answer categories. In the next example, the situation is slightly different, since the title implies recovering information that is absent in the question (underlined in the example).

TITLE: Adequacy of the training provided by the company to do the job
QUESTION: Q13 Have you been provided with information and training to do your job?
-- Yes, enough -- Yes, but insufficient -- No, but I can managed – No and I have difficulties

These examples show quite a complex process of human interpretation of the question and answer categories, like making explicit the implicit, synonymy and paraphrasing. Their automatic processing will call at deep natural language processing techniques, accompanied by domain knowledge, computational lexicons and grammars for natural language understanding and generation. Since we look for a quick and unexpensive solution, deep natural language processing falls out of the scope of this work.

3.1.2 Objective Titles

These titles refer to an objective piece of information about the interviewee. Thus, in their linguistic structure there is not an initial word denoting a judgement but a concrete referent or property. Two types are distinguished: ***specific*** and ***fixed***.

Objective Specific Titles. *They are particular to a given survey and usually refer to habits like smoking, sports, leisure, possession of assets, acknowledgement of persons, etc. They follow the general schema of:*

```
Initial word + nexus + Topic
```

Where initial word can be “Person”, “Entity”, “Possession”, “Likelihood”, “Frequency” … and the topic modifies or characterizes the initial word.

The following is an example of an objective specific title and its question, where it can be observed a clear linguistic relation between both items: part of the title is included in the question (relevant fragments are underlined in the question).

TITLE: Likelihood that Communities will rise, lower, or leave as they are, taxes
QUESTION: Once implemented the new system, and provided that Autonomous Communities can partially modify tax rates of the Income Tax, what do you think is more likely to happen: that communities will raise taxes, lower taxes or leave them as they are?
- Will raise taxes I - Will lower taxes I - Will leave them as they are

The extraction of the relevant pieces of information in this type of titles does not require deep natural language understanding as in the examples of 3.1.1, shallow text processing techniques and even regular expressions will suffice to process them.

Objective Fixed Titles. *They are obligatory in all surveys and refer to the so-called socio demographic variables like Sex, Age, Labour situation or Social class of the interviewee. Apart from any consideration about their linguistic features, titles under this category present two characteristics: they are very frequent and they are exactly repeated over all their occurrences. For example, the following question is exactly repeated 1564 times in the corpus with its exact title:*

TITLE: Age of the Interviewee
QUESTION: How old were you in your last birthday?

So these fixed titles could represent the simplest case of the problem, where a fixed title is assigned to a finite set of questions without further linguistic analysis.

After this preliminary analysis, it is evaluated the amount of questions belonging to the different types present in the corpus of titled questions in order to apply the same percentages to the bulk of untitled questions. It is also interesting to obtain the estimated quantity of titles that result from an interpretation of the whole question (referred as *Non Assignable* titles). To do that, a sample of 240 titles are analyzed, that for a confidence interval of 95%, yields an error rate of 6.3. Estimated frequencies are given in table 1.

Table 1. Frequencies for the different title categories

Title Category	TOTAL	%
Non assignable Titles	59	24,58%
Objective Fixed Titles	89	37,08%
Objective Specific Titles	44	18,33%
Subjective Titles	48	20%

4 Resolution Strategy

This section describes the solution to the problem and how we approach the work in the light of the results of the preliminary analysis. In essence, a title can be viewed as the concatenation of relevant pieces of information that are present in the question (namely, initial word and the topic of the question) and these pieces of information have to be found in the question. For space and clarity reasons, we will have a closer look at objective specific questions in order to illustrate the resolution strategy.

Objective specific questions refer to an objective piece of information about the interviewee addressing a wide variety of topics, including frequencies of actions, possession of things, persons with a give feature or remembering of vote. Due to the variety of topics, the identification of this type of questions relies on a number of linguistic constructions like different configurations for wh-questions mainly (considering wh-questions those headed by the equivalent pronouns of *who*, *what*, *which*, *where*, *when* and *how* in Spanish). Let's have a closer look at two paradigmatic types of objective specific questions: those about frequency of actions and those about persons/entities that do something.

Example 1: Frequency of Actions

Consider the following untitled questions:

> From the following types of alcoholic beverages, could you tell me how often you consume them? (INTERVIEWER: read each type of beverage and SHOW CARD G).

> ONLY FOR THOSE WHO HAVE CONSULTED WITH A PHYSICIAN IN THE LAST TWO WEEKS (1 in Q8). Q9 How many times?

The wh-phrase present in both sentences (*how often* and *how many times*) clearly identifies the intention of the question (*Initial word*) as "Frequency". On the other hand, the *Topic* part of the title is not included in the interrogative sentence. In the first example, the topic is expressed as the pragmatic focus of the text, heading the interrogative sentence. In the second case, the topic is expressed in the interviewers' instruction. This implies that it is required to identify focused extrasentential elements.

The grammar rules that cover these questions are (the rule is adapted to English, although it is originally expressed in Spanish):

```
IF Question =~ /How many times do you <anyWord>
[ObjectPronoun][QuestionMark]/
→ { InitialWord = “Frequency”;
     Content = FocusedTopic}

PROPOSED TITLE: Initial word + “for” + FocusedTopic
```

That is, the presence of an accusative pronoun in the interrogative sentence implies that the Content is expressed outside the sentence, and it triggers the rules for focused topics. The corresponding rule for identifying the focused topic is the following:

```
IF Question =~ /From [ARTICLE] following
<anySequenceOfWords> <PUNC|that>/
→  FocusedTopic = <anySequenceOfWords>
```

Where <anySequenceOfWords> is recognized by means of regular expressions. These two rules generate the title:

TITLE: Frequency of consuming the following types of alcoholic beverages

The second question is similarly processed: the topic is to be found in the interviewer's instruction. So when the interrogative sentence consists of the wh-pronoun and just one word, the topic is extracted from the interviewer's instruction text. Besides, when generating the title, it has to be converted into lowercase characters. The next rule applies:

```
IF Question =~ /ONLY FOR THOSE <WHO|THAT>
<anySequenceOfWords> (QuestionID)* How many times <PUNC>
   → {    Initial Word = “Number of times”
          Topic = lowercase(anySequenceOfWords) }
PROPOSED TITLE: Initial word + “that” + “the interviewee”
+ anySequenceOfWords
```

And it produces the following title:

TITLE: Number of times that the interviewee has consulted with physician in the last two weeks

EXAMPLE 2: “PERSON/ENTITY THAT …”

Within Objective Specific questions, it is frequent to ask about the mere acknowledgement of persons or facts and prejudgement about people. This sort of questions revolve around the pronoun *who* (Sp. *quién*) with variations. Here are some paradigmatic examples:

FROM Q001 and Q013. ONLY FOR THOSE WHO AT PRESENT LIVE AT HOME WITH SOMEONE IN THEIR OWN OR RENTED HOUSING (More than 1 in P001 and 1 or 2 IN P013). Who is the holder of the rental or the owner of this housing? - - (Write down)

```
IF Question  =~ /Who is (AnySequenceOfWords)"?"/
   → {    Initial Word → "Person that"
          Topic → AnySequenceOfWords  }
```

And the following title is generated:

TITLE: Holder of the rental or the owner of this housing

Thus, the general strategy follows a grammar-based approach; where each sentence is subject to the following processes:

1. Extraction of the initial word: identify the linguistic construction that hints it (be it in answer categories or the wh-pronoun).
2. Extraction of the topic
   a. In the interrogative sentence
   b. In the anteposed topic before the interrogative sentence
   c. In the instructions to the interviewer.
3. Generation of the title.
   a. Concatenate both items, add nexus if needed
   b. Include formatting instructions like upper to lower case, substitution of demonstrative of "the", pronoun *Usted* (En. you) for "the interviewee".

5 Results and Conclusions

There are two main aspects to be evaluated: the quantity and the quality of the generated titles. Quantitative results are summarized in table 2. As can be seen, at the end of the process, we were able to generate 22347 titles and leaving apart 1627 questions as filtered ones. This means that we automatically titled around 47% of the questions.

Table 2. Results for untitled questions

Question Status	N
Titled Question	22347
Untitled Question	23990

We also reviewed the quality of the generated titles. To do that, we extracted a sample of 300 titles and evaluated their quality, focusing on two main aspects: legibility of the sentence and presence of relevant information. The average percentage of correct titles for all the samples was **96%.**

Thus after the evaluation, we can ask ourselves again whether our initial hypothesis were correct. From the quantitative point of view, our hypothesis about the frequency of the different types of questions is only partially correct. Untreated questions represented 24% of the titled questions, whereas they represent 50% of untitled questions. Fixed questions are also less numerous in the corpus of untitled questions. However, from a qualitative point of view, we have assured homogeneous and correct titles.

The obtained results made us think about the differences in the distribution of the frequency of the different types of questions. This shift is probably due to the evolution of society that is reflected in the topics of the questions. The followed techniques and strategies also deserve a reflection. Linguistic processing is kept to a minimum, since the linguistic resources are expensive. On the other hand, domain-dependent strategies prove to be highly efficient while quick to be developed.

Acknowledgments. The authors would like to thank Pilar Rey del Castillo for her continuous support and valuable help during the production and evaluation of this work.

References

1. Kupiec, J., Pedersen, J.O., Chen, F.: A trainable document summarizer. In: Proceedings of SIGIR-1995, Seattle, WA, pp. 68–73 (1995)
2. Brandow, R., Mitze, K., Rau, L.F.: Automatic condensation of electronic publications by sentence selection. Information Processing and Management 31(5), 675–686 (1995)
3. Lin, C., Hovy, E.: Identifying topics by position. In: Proceedings of the Fifth Conference on Applied Natural Language Processing (ANLP–1997), pp. 283–290 (1997)
4. Osborne, M.: Using maximum entropy for sentence extraction. In: Proceedings of the Acl-2002 Workshop on Automatic Summarization, vol. 4 (2002)
5. Tseng, Y.-H., Lin, C.-J., Chen, H.-H., Lin, Y.-I.: Toward Generic Title Generation for Clustered Documents. In: Ng, H.T., Leong, M.-K., Kan, M.-Y., Ji, D. (eds.) AIRS 2006. LNCS, vol. 4182, pp. 145–157. Springer, Heidelberg (2006)
6. Kong, S.-y., Wang, C.-c., Kuo, K.-c., Lee, L.-s.: Automatic Title Generation for Chinese Spoken Documents with a delicate scored Viterbi algorithm. In: Spoken Language Technology Workshop, pp. 165–168. IEEE, Los Alamitos (2008)
7. Radev, D.R., McKeown, K.: Generating Natural Language Summaries from Multiple On-Line Sources. Computational Linguistics 24(3), 469–500 (1998)
8. Tucker, R.I., Sparck Jones, K.: Between shallow and deep: An experiment in automatic summarising, Technical Report 632, Computer Laboratory, University of Cambridge (2005)
9. Saggion, H., Lapalme, G.: Generating informative-indicative summaries with SumUM. Computational Linguistics 28(4), 497–526 (2002)
10. Hahn, U., Reimer, U.: Knowledge-based text summarisation: Salience and generalisation for knowledge base abstraction. In: Mani, Maybury (eds.) Advances in Automatic Text Summarisation, pp. 215–222. MIT Press, Cambridge (1999)
11. Spärck Jones, K.: Automatic summarising: The state of the art. Information Processing and Management 43, 1449–1481 (2007)