AUTOMATIC GENERATION OF TITLES FOR A CORPUS OF QUESTIONS

Jesús Cardeñosa, Carolina Gallardo

Abstract: This paper describes the followed methodology to automatically generate titles for a corpus of questions that belong to sociological opinion polls. Titles for questions have a twofold function: (1) they are the input of user searches and (2) they inform about the whole contents of the question and possible answer options. Thus, generation of titles can be considered as a case of automatic summarization. However, the fact that summarization had to be performed over very short texts together with the aforementioned quality conditions imposed on new generated titles led the authors to follow knowledge-rich and domain-dependent strategies for summarization, disregarding the more frequent extractive techniques for summarization.

Keywords: Summarization, text processing, subjective clustering.

ACM Classification Keywords: H.3.1 Content Analysis and Indexing

Conference: The paper is selected from Sixth International Conference on Information Research and Applications – i.Tech 2008, Varna, Bulgaria, June-July 2008

Introduction to the Problem: context and antecedents

Management information based on summaries is frequent in libraries, documentation centers, press or document collections based on short texts. Some libraries have at their disposal information based on summaries of articles or abstracts that have been not generated by the authors but by documentalists. These abstracts have been made in such a way so that information search can be done over the complete document or over the abstract instead. Frequently, these summaries are indexed with some keywords that identify them thematically. Another case is given by news. News are relatively short documents but are identified by means of a title. The title is defined by the author of the new and it serves to search over this news. Titles can be conceived of the most condensed abstraction of the contents of a document. Thus, summaries do not serve a unique function: they can be used for information searches or as indexes of their documents, mainly.

In the area of automatic summarization, summaries can vary in their *form*. Form relates to the way that the summary has been produced, thus the summary can be composed of extracts of the document (extractive summarization) or by and abstract (a concise summary of the central subject matter of a document). Obviously, techniques used for extractive summarization are different from those used for abstracting and more frequent as well. Extractive summarization tries to identify the relevant sentences of the document. To distinguish which sentences are relevant from irrelevant ones, several criteria are used: a positional criterion (e.g., sentences at the end or beginning of the paragraph are considered to be more relevant) was used in [Brandow et al. 95]; [Lin and Hovy, 1997] considered as relevant those sentences that contained signature words (id est, key words that defined by means of frequency measures like the tf-idf schema); on the other hand, [Osborne, 2002] classifies sentences as relevant/non relevant according to the existence of certain word pairs, sentence length, sentence position and discourse features. In essence, extractive summarization typically uses statistical techniques to extract the relevant sentences. As such, they constitute domain-independent techniques.

Non extractive trends in single-document summarization rely on knowledge-based techniques and tend to be domain-dependent approaches. A paradigmatic system, SUMMONS described in [Radev and K. McKeown, 1998], 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 based on templates. Other knowledge-based

approaches make use of linguistic processing, like [Tucker and Sparck Jones, 2005], together with domain knowledge, exemplified by the works of [Saggion and Lapalme, 2002] or [Hahn and Reimer, 1999]. In any case, during the last decade there has been much less research and work in knowledge-based summarization (see [Spärck, 2007] for an up-to-date review of the summarization area).

In our specific case, the problem is defined by the necessity to assign a title to the questions that belong to the opinion polls that the Centre for Sociological Research (CIS – *Centro de investigaciones sociológicas*) of the Spanish government. A question is a short text with the objective of collecting the value or one of more sociological variable. 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. Questions include a number of different answer options, then interviewees have to choose between one or more, free answers are not allowed. Thus, a title for a question is more than an identifier: it is a summary of the content of the question that permits the understanding of the distribution of the frequency of answers without reading the complete text of the question.

Bearing in mind the particular nature of our problem, any extractive technique will prove inadequate for our problem. Since there is a need to generate new titles that are a condensed abstracting of the question, the strategies to follow will have to be knowledge-rich and domain-dependent. This article will describe the process that we have followed to automatize the process of title generation. Although our proposal is based on domain-dependent criteria, it is applicable to similar problems.

Analysis of the Domain

The specific problem is as follows. The complete corpus of different questions belonging to the opinions polls is composed of:

- 39257 questions with a title, assigned manually by experts.
- 47964 questions with no title (hereafter, they will be referred as untitled questions).

Since there is a corpus of already titled questions, the generated titles have to be coherent and similar with the manually assigned ones. To do that, we performed a thorough analysis of the domain with the main objective of searching for relations (of any kind) between the existent questions and their titles and extrapolate such relations to the corpus of untitled questions. Although there are many aspects that can be analyzed in this particular problem, we focused on two main aspects: linguistic features of titles and relations between titles and questions. The analysis is completed with the estimation of frequencies of the different types of titles and relations.

I. Linguistic features of titles

The set of titles (without questions) were thoroughly examined. Aspects like length of the title, existence of repeated titles, linguistic constructions and tackled themes were taken into account. For our purposes, we established two broad categories of titles: subjective topic titles and objective specific titles.

Subjective topic titles

This sort of titles refers to the judgement about a given topic of the interviee. The judgement can be an approval, rejection, preference, evaluation, etc. of a topic; and this sort of judgement is included in the title, together with the object of the judgement. It is remarkable that from a thematic point of view, the most frequent tackled themes are vote, elections and politicians. Besides, the titles dealing with these three topics shows less variation that others. In general, the structure of subjective topic titles follow the general schema of: *Type of judgement + Nexus + Topic.* Where type of judgement is a word like "opinion", "preference", etc., "nexus" is a the preposition that requires the initial word, and topic is the nominal group denoting what the question is about. Table 1 shows some examples.

Table 1. Example of Subjective topic titles

Type of Judgement	Nexus	Topic
Approval	of	the labour of Felipe Gonzalez as Prime Minister
Preference	between	different alternatives of territorial organization of the State
Satisfaction	with	the job of the interviees that do not study

Objective Specific Titles

These titles asked for an objective piece of information about the interviee. Thus, in their linguistic structure there is not an initial word denoting a judgement. There are two types:

- *Fixed*: they seem to be obligatory in all surveys and usually refer to sociodemagraphic aspects of interviees like labour situation, social class, etc. Some examples are:
  - Age of Interviewee; Head of the familiy
- *Specific*: they are specific to a given survey and usually refer to habits like smoking, sports, leisure, etc. For instance:
  - Starting Age for smoking; Number of mobile phones of the interviewee

II. Linguistic relation between titles and questions

The degree of success in automatizing the process of generating titles is directly related with the relation between a question and its title. We have identified three main types of relations:

a) The title is a summary of the question. As in the following example:

TITLE: Acceptance of fraudulent behaviour regarding the National Institute of Employment
QUESTION: In our society, there are happen things that are completely acceptable for some people and absolutely unacceptable for others. I am going to read you some of those things and I'd like to know whether they are acceptable or not for you. [….]
- To evade taxes | - To receive an unemployment subsidy while working. | ….

b) The topic of the title is a nominalization or paraphrasis of a part of the question.

TITLE: Frequency of attendance to religious services
QUESTION: How often do you go to mass or attend religious services if you have other religion?
- Never | - Several times a year | - Sometimes in the month

c) The topic of the title is a literal fragment of the question.

TITLE: Attitude towards the creation of an International Court
QUESTION: In any case, are you in favour or against the creation of an International Court of these characteristics.
- In favour | - Against | - Not know – not answer

III. Frequency of the different types of titles & questions

Our work is based on the hypothesis that the analysis results obtained from the existent corpus of titled questions are valid for the corpus of untitled questions. In this way, we assure similarity of the generated titles with the existent ones. It is important to estimate the frequency of the different types of titles and of the different types of relations between titles and questions. To estimate frequencies, we established five categories of questions that resulted from the reorganization of the classification from the analysis of the linguistic features of titles and from the relation between title and question, namely:

- From the linguistic analysis of titles, there results four categories. *Subjective topic* titles branches into two categories: a) Topics about vote, elections and politicians; and b) Rest of topics. Whereas *Fixed Objective Specific* titles divides into other two categories: c) Fixed; and d) Specific.
- From the relation between title and question, we are just interested in identifying the untreatable questions, since these are the ones that will pose more problems to the task. Thus, only one category is posed: Untreatable questions.

Thus, we will postulate 5 categories: (1) Topics questions about vote, elections and politicians; (2) Rest of topics questions; (3) Fixed question; (4) Specific questions; and (5) Untreateble questions. In order to estimate the frequency of these five categories, we extracted six samples of 40 questions from the corpus of titled questions. For each sample, we classified the question –along with its title- under one of the five postulated categories. Table 2 shows the estimated frequency of different types of titles.

Table 2. Frequency of different types of titles

Type of Question	Frequency
Elections, vote and politicians	14.58 %
Topic questions	12,92%
Specific questions	10.83 %
Fixed questions.	37,08 %
Untreatable questions	24,58 %

Under the hypothesis that untitled questions behave as titled ones, these frequencies helped us to estimate that around 25% of the corpus of untitled questions would remained untitled.

Development of the work: Methodology

We imposed the following working hypothesis: only questions whose title can be generated from a nominalization, paraphrasis or exact wording of a fragment of the question will be assigned a title. We will not include deep NLP processing technology but shallow language analysis and domain-dependent patterns to identify the relevant fragments of the sentence and to generate the corresponding title.

The main reason to reject deep natural understanding techniques are given by the number of language resources required, such us dictionaries, grammars for analysis and grammars for generation. A domain-dependent strategy instead does not require either deep natural language understanding or big language resources. On the other hand, it requires an exhaustive domain-analysis and a representative corpus of examples.

Let's examine the following example:

TITLE: Religiosity of the interviewee
QUESTION: How would you considered yourself regarding religion?
- Practising catholic. | - Non practising catholic | - Other religions | - Non believer | - Do not answer

This specific question and its title are repeated around 800 times in the corpus of titled questions however it will fall into the category of untreatable questions. It represents a simpler case where to fairly repeated string it is assigned a fixed title, disregarding any kind of linguistic analysis.

Thus the chosen technique was a domain-dependent one. This option was supported by the following facts: a) high frequency of fixed questions; b) homogeneity in the existent titles; and c) the results of the analysis of the domain already pointed out at patterns.

Strategies for the different types of questions

Each type of questions have a different strategy. For example, fixed questions are the simplest, their treatment imply looking for a specific string and assign another string, while the rest of categories require to look for the topic, type of opinion, etc. Let's have a look at how each different type have been dealt with.

I. Questions about vote, elections and politicians

The questions about these topics fairly frequent in the corpus of both titled and untitled questions. They show little variability in their wording, being the variable elements the type of election (European, general, regional or municipal), the date of the election or the politician that is being evaluated.

The strategy for generating the titles for this type of question is a) to associate a specific sequence of words in the question to a specific partial title; b) identify the variable elements (type and date of election, name and position of the politician); and c) concatenate the different elements. That is, these titles have the following general schema:

TITLE → Partial title + Variable element

For example, pattern 1 shows one specific pattern for the identification of a partial title.

PATTERN 1. PARTIAL TITLE: "INTENTION OF VOTE IN …"

```
[Q]/Do you think you will vote in/ → [PT] "Intention of vote in"
```

That is, if the string "do you thing you will vote in" matches the question [Q], then identify the partial title [PT] as "Intention of vote in".

Pattern 2 identifies one variable element (type of election).

PATTERN 2. VARIABLE ELEMENT: TYPE OF ELECTION

```
[Q] /<these|the|forthcoming> elections X PUNCTUATION/ → [VE] "the X elections"
```

In this case, if the question matches any of the words <these|the|forthcoming> followed by the string "elections" and any string (denoted by X) until a punctuation mark, then identify the variable element [VE] as the string "elections X".

The following question and title illustrate these patterns:

QUESTION: Do you think you will vote in the elections to the European Parliament, to be celebrated next 15th June.
TITLE: Intention of vote in the elections to the European Parliament.

II. Rest of topic Questions

The strategy to generate the titles for these questions follows two steps: a) identify the sort of judgement that is being required to the interviee (that is, an opinion, approval, reason or evaluation); and b) identify the topic of the question. The final title will be the concatenation of the type of judgment and the topic:

TITLE → Type of judgment + topic

Both elements are extracted with the help of regular expressions, but in some cases some linguistic knowledge is also used. Pattern 3 shows an specific example to identify the type of judgment, whereas pattern 4 identifies the topic of a question.

PATTERN 3. JUDGEMENT TYPE: "DEGREE OF AGREEMIENT WITH"

```
[Q] /TELL <the|your> degree of agreement/ → [JT] "Degree of agreement with"
```

That is, if the question is composed of any form of the verb "to tell" followed by determiner "the" or pronoun "you", followed by the string "degree of agreement with", identifies the judgement type as "degree of agreement with".

PATTERN 4. TOPIC

```
[Q] /going to read (you) some <opinions | statements> about X FINAL_MARK/ → [TOPIC]: "some opinions about X"
```

That is, if the question matches the string "going to read you some opinions about" and its possible variations (presence or absence of "you", disjunction between "opinions" or "statements") followed by any string (represented by X) until a final mark, identify the topic of the sentence as "some opinions about X".

FINAL MARK

```
[FINAL_MARK]: <;|:|.|CLAUSE>
```

That is, a final mark can be any of the following: a semi-colon, a full stop, colon or the beginning of a new clause (for example: pronoun + verb).

These patterns show how (very) shallow linguistic knowledge aids to the task of identifying patterns. They can be directly applied to the following question, where both the type of judgement and topic are highlighted:

QUESTION: Now, I am going to read **some opinions about the development of the State of Autonomies** and I would like you to tell me your **degree of agreement with** each of them. Autonomous Communities …
- Have contributed to …

TITLE: "degree of agreement with the development of the State of Autonomies".

III. Specific Questions

These questions usually ask for objective data about the interviee. Thus they present more variability in the topics of the questions. For this reason, the strategy is slightly different from the previous ones: patterns for these questions do not rely on the identification of a particular string in the sentence but on the identification of specific linguistic constructions. Thus, for the process to be successful, it is required that the linguistic constructions present in the questions are homogeneous. Consequently, this type of questions are the ones that require more linguistic knowledge for their processing, in particular, the recognition of the shallow structure of wh- questions. Patterns 5 and 6 are two paradigmatic examples for specific questions.

PATTERN 5. TITLE: "NUMBER OF TIMES …"

`[Q]: /How many times (in total) have you` **X**`?/ → [T] "Number of times that the interviee has` **X"**

That is, if the question matches the string "how many times have you" followed by any string (represented by "X") and a question mark, identify the title [T] of the question as "Number of times that the interviee has X".

This pattern has been directly applied to the following question:

QUESTION: How many times have you been hospitalized in the last twelve months?
TITLE: Number of times that the interviee has been hospitalized in the last twelve months.

PATTERN 6. TITLE: "PERSON/ENTITY THAT …"

`[Q]: /Who do you <think|believe> that` **X** `<:|?>/ → [T] "Person/Entity that` **X"**

That is, if the question matches the string "who do you think that" (or "believe", instead) followed by any string (represented by "X") and a question mark, identify the title [T] of the question as "Number of times that the interviee has X".

Pattern 6 generates the following title:

QUESTION: Who do you think that should provide information about the social and sanitary assistance and services for old people?
TITLE: Person or entity that should provide information about the social and sanitary assistance and services for old people

IV. Fixed Questions

The strategy for the generation of this titles does not require any linguistic knowledge. The process merely consists in the assignment of a fixed title (without variations) to questions that present a particular wording with hardly any variation. Patterns 7 and 8 deal with two different types of fixed questions.

PATTERN 7. FIXED TITLE

`[Q]: /what is your social class/ → [T]: "Subjective social class of the interviee"`

That is, if the question matches the string "what is your social class", identify its title as "Subjective social class of the interviee"

PATTERN 8. FIXED TITLE

`[Q]: /Which of the following describes your current situation?/ → [T] "Labour situation of the interviee"`

That is, if the question matches the string "Which of the following describes your current situation?", identify its title as "Labour situation of the interviee".

V. Untreatable questions

These questions do not have any feature to identify them. In this case, we do not attempt to generate a title.

Finally, in addition to untreatable questions, there is a set of questions that are left untitled intentionally. These questions does not have enough content to generate a title. In particular, they present any of the following characteristics:

a) The question has less that 4 words. For example:

QUESTION: And why not?

b) The question ends with suspensions dots. For example:

QUESTION: What do you think about …?

c) The questions contains enclitic pronouns and general terms like "that", "statement", "reason" in their topic. For example:

QUESTION: Do you agree with that?

Thus, any question with any of the aforementioned characteristics is left out from the process of title generation. They are referred to as filtered questions.

Results

There are two main aspects to be evaluated: the quantity and the quality of the generated titles. Quantitative results are summarized in table 3. 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 3. Results for untitled questions

Type of Question	Number
Titled questions	22347
Untitled questions	23990
Filtered Questions	1627
TOTAL	47964

We also reviewed the quality of the generated titles. To do that, we evaluate the quality of titles of six samples of 50 titles each. 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.

Conclusions

There are several conclusions from this work. The first one made us think about the differences in the distribution of the frequency of the different types of questions. This shift is sometimes due to the evolution of society. The sociological changes are reflected in the topics of the questions.

The second one refers to the followed techniques and strategies. In this work, domain-analysis and ad hoc patterns prevails over domain-independent linguistic processing. 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.

Bibliography

Brandow, R., Mitze, K., and Rau, L. F. (1995). Automatic condensation of electronic publications by sentence selection. Information Processing and management, 31(5), 675–686

Hahn, U., and Reimer, U. (1999). 'Knowledge-based text summarisation: Salience and generalisation for knowledge base abstraction'. In Mani and Maybury, eds. (1999). Advances Advances in automatic text summarisation. Cambridge, MA: MIT Press. (pp. 215–222).

Lin, C., and Hovy, E. (1997). Identifying topics by position. In Proceedings of the Fifth Conference on Applied Natural Language Processing (ANLP--97), 283--290.

Osborne, M. (2002). Using maximum entropy for sentence extraction. In Proceedings of the Acl-02 Workshop on Automatic Summarization - Volume 4

Radev, D. R. and McKeown, K. (1998). Generating Natural Language Summaries from Multiple On-Line Sources. Computational Linguistics, vol 24(3): 469-500.

Saggion, H. and Lapalme, G. (2002). Generating informative-indicative summaries with SumUM. Computational Linguistics, 28(4), 497–526.

Spärck Jones, K. (2007) Automatic summarising: The state of the art. Information Processing and Management, 43: 449-1481

Tucker, R. I. and Sparck Jones, K. (2005). Between shallow and deep: An experiment in automatic summarising, Technical Report 632, Computer Laboratory, University of Cambridge.

Authors' Information

Jesús Cardeñosa – Group of Validation and Industrial Applications. Facultad de Informática. Universidad Politécnica de Madrid; Madrid 28660, Spain; e-mail: carde@opera.dia.fi.upm.es. http://www.vai.dia.fi.upm.es

Carolina Gallardo – Group of Validation and Industrial Applications. Escuela Universitaria de Informática. Universidad Politécnica de Madrid. Carretera de Valencia Km.7. 28041 Madrid; email: cgallardo@eui.upm.es. http://www.vai.dia.fi.upm.es.