Document Title Patterns in Iinformation Retrieval *

Manuel Montes-y-Gómez[1], Alexander F. Gelbukh[1], and Aurelio López-López[2]

[1] Natural Language Laboratory,
Center for Computing Research (CIC), National Polytechnic Institute (IPN),
Av. Juan de Dios Bátiz, CP 07738, Zacatenco, Mexico City, Mexico
mmontesg@susu.inaoep.mx, gelbukh@pollux.cic.ipn.mx
[2] INAOE, Electronics.
Luis Enrique Erro, No. 1, Tonantzintla, Puebla, 72840, Mexico
allopez@gisc1.inaoep.mx

Abstract. The document titles give an important information about documents. This is why they are frequently used to obtain document keywords. We use them to determine document intentions. To obtain some textual details, we use special information extraction techniques for the construction of extra-topical representations of the documents. This representation reflects a document more completely. A possible use for the representation in the information retrieval is described. This improves the retrieval results.

1 Introduction

Unlike the structured information or formal representations, raw texts have a very complex form. This allows them to describe more completely all entities and facts, but at the same time provokes many of the difficulties in the analysis.

Nowadays, almost every raw text operation, for example, text classification, information retrieval, text indexing or description, is done on the basis of keywords or, in the best case, of topics obtained from entire texts of some their parts. This method can give text characteristics beyond topicality, such as intentions, proposes, plans, etc., which are usually ignored [1].

In this paper, we reveal the link between document title and its author intentions. We also describe a method of the automatic extraction of the intentions and finally propose a possible use of this information in IR systems.

2 Intention Structure

By intention, we mean determination to do something. In this sense, intentions are related with some acts fixed in the document text. They are grammatically

* This work was done while Manuel Montes y Gómez was supported by CONACYT, Mexico, through scholarship to pursue his Master Sc. Degree. The work was also partially supported by REDII-CONACYT and DEPI-IPN, Mexico.

associated with some verbs having the main topic of the document as their subjects, such as *introduce, describe* or *propose*.

On the basis of these features, the task of determining document intentions consists of finding verbs which actions are performed by the document. For instance, the intention of some document is to *describe* something if there is some evidence in the document body that relate the document with the action *to describe*.

With this approach, the extraction of the document intention it is not simple. Intentions are more than mere actions reflected, they additionally include an object of the action and sometimes more pieces of related information. For instance, it is not sufficient to say that the intention of some document is to describe. It is also necessary to indicate what is to be described (the object), as well as how, when, or why this action is done.

3 Title Patterns

A title is the part of the document most heavily used for such tasks as indexing and classification. Just this prompts us to use titles for extraction of the intentions. We can note the following facts about the relation between titles and intentions [4]:

- Intentions are associated with a noun pattern:
  - A noun is followed by a preposition *of* or *to* in the beginning of the title (*An Introduction to a Machine-Independent Data Division*)
  - A substantive coordinated group is followed by a preposition *of* or *to* (*Implementation, evaluation, and refinement of manual SDI service*)
  - The case is similar to the previous, but with a dash instead of conjunction (*Computer simulation - discussion of the technique...*)
- Intentions are related to some gerund patterns:
  - The gerund is at the beginning of the title (*Proving theorems by recognition*)
  - The sequence adjective - gerund starts the title (*Automatic indexing and generation of classification systems*)
  - Prepositional group with gerund is anywhere except at the end (*A language for modeling and simulating dynamic systems*)

4 Intention Extraction Method

The system of intention extraction developed by us follows a information extraction scheme [2]. It contains a tagger, a filtering component, a parser, and a module of generation of the output data. As an example, let us process the title *Algebraic Formulation of Flow Diagrams*. First, each word is supplied with a syntactic-role tag[3]:

[3] The Tagger we are using is based on the Penn Treebank Tagset.

Algebraic|JJ formulation|NN of|IN flow|NN diagrams|NNS|

The next component selects only the titles containing some information about intentions. This filtering is based on the patterns previously described. Then the chosen titles are parsed and their structured representation[4] is formed [5]:

[[np,[n,[formulation,sg]],[adj,[algebraic]],
[of,[np,[n,[diagram,pl]],[n_pos,[np,[n,[flow,sg]]]]]]],'.']

This representation is entered to the last component, i.e., the output generator, where the structured representation is transformed into a semantic representation of the document (a conceptual graph) [3]:

[formulate] > *(manr)* > *[algebraically]*
[formulate] > *(obj)* > *[flow-diagram,{*}]*

5 Experimental Results

Our system was tested on a collection of 4663 documents. Manual evaluations gave 802 useful documents (17.2% of their total number), while our system 738 (15.7%). The low percentage of the documents that can be processed by our method does not mean its low usefulness. The described method of intention extraction from titles is to be used together with our method of intention extraction from abstracts [1], [4]. As it is shown below, the two methods work on nearly complementary distributed sets of documents and together cover up to 90% of the collection.

Analyzing only those documents from the collection that have abstracts (normal situation), we obtained the following results:

Table 1. Abstracts and Titles

	Number	Procents
Documents with abstract	1587	100 %
Documents with graph from abstract	1207	76 %
Documents with graph from title	301	19 %
Documents with at least one graph	1438	**90 %**

Thus, for collections of documents with abstracts, the percentage of documents that are assigned at least one conceptual graph is close to 90%, being sufficient for any retrieval application.

[4] The parser we are using was created in the New York University by Tomek Strzalkowski. It is based on "The Linguist String Proyect (LSP) Grammar" designed by Naomi Sager.

6 Future Infomation Retrieval Application

With the electronic information explosion caused by the Internet, more and more diverse information is on hand, so the need in better search engines is staggering. The more information about documents we have, the better we can evaluate the documents.

Basing on these ideas, we designed a new IR system. It will perform the document selection taking into account two different levels of document representation. The first is the keyword document representation. On this level, the documents are represented by means of keywords and the search is done by traditional retrieval method. The first level will select all documents related to the given general topics.

The second level is formed with semantic representations of the document intentions. This second level complements the topical information about documents and provides a new way to evaluate the relevance of a document. Intentions of documents are extracted from titles and abstracts [1], [6].

7 Conclusions

With this article and with [1], [4], [6], we try to break down the keyword representation paradigm and begin to use other document characteristics. This paper shows the relations between document titles and their intentions and demonstrates how these intentions are reflected in titles. The method of intention extraction is domain independent, so that can be applied to documents on any topic.

References

1. López-López, A. and Myaeng, S. H.: Extending the Capabilities of Retrieval Systems by a Two Level Representation of Content. In: Proc. of the 1st Australian Doc. Comp. Symposium (1995)
2. Cowle, J. and Lehnert, W: Information Extraction. Com. of the ACM, 39 (1), (1996)
3. Sowa , J. F.: Conceptual Structures: Information Processing in Mind and Machine. Addison-Wesley (1983)
4. Montes y Gómez, M.: Extracción de Información de Títulos de Documentos. M.Sc. Thesis, Electronics, INAOE, México (1998)
5. Strzalkowski, T.: TTP: A fast and Robust Parser for Natural Language.PROTEUS. Project memorandum 43-A (1992)
6. López-López., A. and Montes y Gómez, M.: Nominalization in Titles: A Way to Extract Document Details. In: Memoria del Simposium Internacional de Computación CIC'98, México (1998)

This article was processed using the LaTeX macro package with LLNCS style