Published in Procs. IJCAI 99 Workshop IRF-3: Text Mining: Foundations, Techniques and Applications, 31 July-6 August, Stockholm, Sweden, pp. 101-102.

Extraction of Document Intentions from Titles

Manuel Montes-y-Gómez,
Alexander F. Gelbukh

mmontesg@susu.inaoep.mx
gelbukh@pollux.cic.ipn.mx
Natural Language Laboratory,
Center for Computing Research (CIC),
National Polytechnic Institute (IPN).
Av. Juan Dios Bátiz, Zacatenco, 07738 D.F.
Mexico

Aurelio López-López

allopez@inaoep.mx
Electronics, INAOE
Luis Enrique Erro No. 1
Tonatzintla Pue. 72840
México

Abstract

In spite of their small size, titles are a very important source of information about document contents. This is why they are frequently used as a way to obtain document keywords and this is the reason we have chosen to use them to obtain and extract document intentions. In order to construct better document representations, we analyze the opportunities to extract some document details from titles. Particularly, we propose to use some classical information extraction techniques for constructing extratopical representation of the documents. It is put together with the keywords to form a new and more complete representation of the document. A possible use for this representation in the Information Retrieval area is described, as well as how this paradigm for document representation can improve the actual retrieval results.

1 Introduction*

Unlike the structured information or formal representations, raw texts have very free and complex form. These characteristics allow them to describe better and more completely all entities and facts, but at the same time these features provoke many analysis difficulties.

Nowadays, almost every raw text operation, for example, text classification, information retrieval, indexing or text description, is done on the basis of keywords or, in the best case, of topics or themes obtained from some parts of the documents or from the entire text [Guzmán, 1998]. This paradigm generally leads to ignoring text characteristics beyond topicality, such as intentions, proposes, plans, content level, etc. [López-López and Myaeng, 1997].

In this paper, we present evidence of the relationship between document intentions and document title. Additionally, we describe a method used for automatic extraction of the document intention(s) and finally proposed a possible use for this information in IR systems.

2 Intention Structure

By *intentions*, we mean determination to do something. Intentions describe, or are related with, the act intended by the document. They are grammatically associated with some verbs that take the main topic of the document as their subject, such as *introduce*, *describe*, *propose*, etc.

The task of determining the document intentions consists in finding verbs whose actions are performed by the document. For instance, we can say that the intention of some document is to *describe* something if there is some evidence in the document that relates the document with the action "describing."

With this approach, extraction of the document intention might seem a simple task; in fact, it is not. Document intentions are more than simple actions related with the documents. They include an action, an object of the action, and sometimes one or more pieces of related information.

For instance, it is not sufficient to say that the intention of some document is to describe, since it is also necessary to indicate what thing the document describes (the object), and it can be necessary to say how, when or why this action is done (some related information).

3 Intentions in Document Titles

Title is not only the very first information the reader receives from a document, but also the part of the document most heavily used for such tasks as indexing and classification. This background inspires us to use titles for extraction of the intentions, We can note the follow-

* The work of M. Montes-y-Gómez was supported by CONACyT scolarship, México. A. Gelbukh was partially supported by REDII-CONACyT, DEPI-IPN, and SNI, México.

ing facts about the relation between titles and intentions [Montes-y-Gómez, 1998]:

- Document intentions are associated with title nominalizations, e.g.: "Numerical solution of the polynomial equation, " "*An Introduction to a Machine-Independent Data Division.*"
- Intentions are also related to some present participle patterns, e.g.: "*Proving theorems by recognition,*" "Computing radiation integrals."

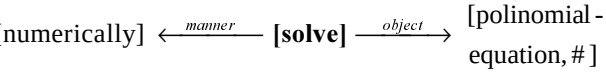
4 Intention Extraction Method

The intention extraction system we developed follows a classic information extraction scheme [Cowie and Lehnert, 1996]. It consists of a tagger, a filtering component, a parser, and a module of generation of output data.

5 Output Representation

The output representation generated by our system is a Conceptual Graph [Sowa, 1983]. This representation is a network of concept nodes and relation nodes, where concept nodes represent entities, attributes, or events, while relation nodes identify the kind of relationship that holds between the concept nodes.

These graphs make it easy to represent the information about the document intentions. This representation permits to easily use this information in many applications. The following graph illustrates the intention of the document (in boldface), and its structure with the additional information, for the first title given above.



6 Experimental Results

We tested our system on two standard test document collections (CACM-3204 and CISI-1459), consisting of a total of 4663 document descriptions. When comparing the extraction effectiveness against manually identified titles, the method produced a recall and precision of 92%, 98% and 90%, 96%, for CACM and CISI respectively. When analyzing how the intentions from titles complement those identified from abstracts, we achieved as much as 90% of the documents with some kind of intention representation.

7 Conclusions

With this and related works [López-López and Myaeng, 1997; López-López and Tapia-Melchor, 1998; Montes-y-Gómez, 1998], we try to break down the keyword document representation paradigm and begin to use other document characteristics in their representations. In particular, this paper provides evidence for the relations between document titles and their intentions and also demonstrates that these intentions are reflected in titles by some particular nominalizations and present participles patterns. As one of its principal features, the automatic intention extraction method has domain independence, so that it can be applied to documents on any topic.

At this moment we are implementing a new IR system, using a new representation of documents – two-level representation, aimed to improve the information retrieval results, mainly so-called normalized precision. In the future, we plan to apply similar methods to other parts of the documents and develop a better content-detail representation.

References

[Chisnell, *et al*, 1993] Ch. Chisnell, D.V. Rama and P. Srinivasan. Structured Representation of Empirical Information. In *Case-based Reasoning and Information Retrieval,* Technical Report 55-93-07, AAAI Press, 1993.

[Cowie and Lehnert, 1996] Jim Cowie and Wendy Lehnert. Information Extraction. *Communications of the ACM*, 39(1):80-91, January 1996.

[Guzmán, 1998] Adolfo Guzmán Arenas. Finding the Main Themes in a Spanish Document. *Expert Systems with Applications*, 14:139-148, 1998.

[López-López and Myaeng, 1997] A. López-López, and Sung H. Myaeng. Extending the Capabilities of Retrieval Systems by a Two Level Representation of Content. In *Proceedings 1st Australian Document Computing Symposium,* Part I, pages 15-20, Melbourne, Australia, March 1996.

[López-López and Montes-y-Gómez, 1998] A. López-López, and M. Montes-y-Gómez. Nominalization in Titles: A Way to Extract Document Details. In *Proceedings of the Simposium Internacional de Computación CIC'98,* pages 396-404, México D.F. November 1998.

[López-López and Tapia-Melchor, 1998] A. López-López, and Ma. del P. Tapia-Melchor. Automatic Information Extraction from Documents in WWW. In *Proceedings of the VIII International Congress of Electronics, Communications, and Computers CONIELECOMP 98,* pages 287-291, Cholula, Puebla, México. Feb.1998.

[Montes-y-Gómez, 1998] M. Montes-y-Gómez. Information Extraction from Document Titles, M. Sc. Thesis, Electronics, INAOE, México, 1998.

[Sowa, 1983] John F. Sowa. *Conceptual Structures: Information Processing in Mind and Machine*. Addison Wesley, 1983.