Chapter 23
Language Independent Summarization Approaches

Firas Hmida
LINA Nantes-University, France

ABSTRACT

In this chapter, the authors introduce monolingual and multilingual summarization and present the problem of dependence of language and linguistic knowledge of the process. Then they describe the most influential works and techniques in the field of automatic multilingual and language-independent summarization. This section is presented as a solution to solve the problem already explained. The authors present several language independent approaches and used techniques. In the next section, they study the behavior of these methods by discussing their limitations and perspectives.

INTRODUCTION

Automatic summarization is a difficult problem of Natural Language Processing (NLP). It is highly language dependent. Indeed, the increasing volume and variety of electronic information, whether on the Internet or networks companies, makes it difficult to access the required information without recourse to language tools. This makes it very complicated to generate summaries because it requires language skills and knowledge of the world that remain virtually impossible to incorporate in a computer system.

Since multilingual summarization stems from the monolingual summarization, the former have to face many language-dependent challenges in order to be able to deal with different languages. To avoid this problem a lot of works in this area have turned towards language independent summarization.

In the first section we describe the task of summarization and present the problem of dependence of language and linguistic knowledge of the process. Then we describe the most influential work and techniques in the field of automatic multilingual summarization. This section will be presented as a solution to solve the problem already explained in the introduction. We present several language independent approaches and used techniques. In the next section we study the behavior of these methods by discussing their limitations and perspectives. As usual, we will end this chapter with a conclusion referring to the automatic evaluation of multilingual summaries.

DOI: 10.4018/978-1-4666-6042-7.ch023

BACKGROUND

An automatic summarization synthesizes a compressed representation of an information source while maintaining the important information of the original content. It is a very complicated task. However, generally, people still produce summaries so efficiently. Works in this field aimed to imitate the cognitive process of generating a summary. Since a long time, researches have focused on scientific documents and also on press reports. This work deals only with text summarization. We can distinguish two types of summaries: the first one is the single-document summary, when the source document is unique, whereas, in the second one, the multi-document summary, analyzed information may come from several documents. The summary can also have different purposes: It can be generic if it treats all the topics in a document with the same degree of importance, but if it deals with only one specific part of the information required, it is called an oriented summary.

One can think of an approach to summarization as being an extract or an abstract method, with rather different implications. The method using the extraction consists on selecting textual units (words, sentences, etc…) which are supposed to contain important information from the document and then assemble those units to produce an "extract." In other words, an extract is a part taken from a source document in order to provide an overview of its content (Boudin, 2008). An "abstract" is to understand the contents of a source document and reformulate them. It is a gloss describing those contents with an implicit way, which means that they don't have to feature with the same language used in the original document. (Lin and Hovy, 2003) said that nearly 65% of the sentences in manually created summaries are extracted from the source document without any modification.

The multilingual summarization stems from the monolingual automatic summarization: They both have the same functionalities, but the multilingual summarization comes up with a new dimension: globally, it is defined as a process that involves more than one language in the automatically text summarization process.

Multilingual/language-independent summarizer needs to show a good and an equal ability to deal with different languages without any special adjustment such as modifying the algorithm and/or requiring some additional data in each language. The majority of multilingual summarization of documents/queries, written in a foreign language, use automatic (and/or manual) translation as a pre-processing or/and a post-processing step (Chenn & Lin, 2000) and (Ogden, & al.1999). In such cases, the summarization accuracy may be considerably affected by the quality of translation. The automatic translation is known as a very complicated and challenging process, and existing tools usually suffer from the lack of the results' accuracy, therefore, summarization systems are facing a noisy output. In order to deal with multilingual contents, only few works do not employ translation or any other complementary tool. Another simple technique has been included by the authors of (Salton & al. 1997), (Radev & al.2004); it consists in using a graph representation of the text and a similarity measure between the text units that can be easily applied to several languages.

TOWARDS LANGUAGE-INDEPENDENT SUMMARIZATION

The following list is a summary of some features that we need to take into consideration when dealing with a multilingual summarization system. Particularly, these are challenges we must reveal when working with multiple language.

- **Tokenization:** Because languages encode word boundaries differently, tokenization is a first obstacle to overcome when