Extractive Summarization Based on Word Information and Sentence Position

Carlos MÉNDEZ Cruz and Alfonso MEDINA Urrea

GIL IINGEN UNAM
Apartado Postal 70-472, 04510 Coyoacán, DF, MEXICO
{cmendezc,amedinau}@iingen.unam.mx

Abstract. This paper describes an unsupervised experiment of automatic summarization. The idea is to rate each sentence of a document according to the information content of its graphical words. Also, as a minimal measure of document structure, we added a sentence position coefficient.

1 Introduction

Plenty of research has been conducted in the field of automatic summarization. The need for such methods has motivated the exploration of many approaches which reflect the field's complexity. Many aspects of texts must be considered, such as word frequency, document, paragraph and sentence structure, topic and focus structure, information content, etc. Many of these approaches are based on the idea that the greater number of times a linguistic structure or part of it occurs (a word, phrase, sentence, etc.), the more attention the reader will pay to it except when it is a function or grammatical word. That is, a document's sentences receive different levels of attention by human readers. Current interests among researchers are multi-document summarization [1, 2], the application of artificial intelligence methods such as genetic algorithms [2], the use of lexical chains and web resources such as WordNet [3].

Since we are currently developing an open, Spanish language corpus on engineering (CLI) to be available on the Internet [4], we are exploring some summarization techniques to apply to it. The main criteria for this very first experiment was to avoid the heavy techniques that have been and can be developed if one takes into account the complexity of document, paragraph, sentence, and word structure. Thus, we opted for an unsupervised approach based on simple information content measurements that could conceivably be applied to other languages. Actually, information content estimates are typically used for a wide variety of unsupervised tasks. And in fact, some experiments have explored the notions of information content and entropy models for some aspect or another of automatic summarization — for instance, summary evaluation or reductive transformation [5–7]. In this paper, we will first define some basic concepts. Then, we will briefly describe our application and lastly, we will present results and evaluation strategy.