Assessing the Impact of Lexical Chain Scoring Methods and Sentence Extraction Schemes on Summarization

William Doran, Nicola Stokes, Joe Carthy, John Dunnion.

Department of Computer Science,
University College Dublin, Ireland.
{William.Doran, Nicola.Stokes,Joe.Carthy,
John.Dunnion}@ucd.ie

Abstract. We present a comparative study of lexical chain-based summarisation techniques. The aim of this paper is to highlight the effect of lexical chain scoring metrics and sentence extraction techniques on summary generation. We present our own lexical chain-based summarisation system and compare it to other chain-based summarisation systems. We also compare the chain scoring and extraction techniques of our system to those of several other baseline systems, including a random summarizer and one based on tf.idf statistics. We use a task-orientated summarisation evaluation scheme that determines summary quality based on TDT story link detection performance.

1 Introduction

Summarisation is a reductive transformation of a source text into a summary text by extraction or generation [13]. It is generally agreed that automating the summarisation procedure should be based on text understanding that mimics the cognitive processes of humans. However, this is a sub-problem of Natural Language Processing (NLP) and is a very difficult problem to solve at present. It may take some time to reach a level where machines can fully understand documents, in the interim we must utilise other properties of text, such as lexical cohesion analysis, that do not rely on full comprehension of the text.

Lexical cohesion is the textual property responsible for making the sentences of a text seem to "hang together", indicated by the use of semantically related vocabulary [10]. Cohesion is thus a surface indicator of the discourse structure of a document. One method of representing this type of discourse structure is through the use of a linguistic technique called lexical chaining. Lexical chains are defined as clusters of semantically related words. For example, {*house, loft, home, cabin*} is a chain, where *house* and *home* are synonyms, *attic* is part of a *house* and *cabin* is a specialisation of *house*. The lexical chaining algorithms discussed in this paper identify such lexical cohesive relationships between words using the WordNet taxonomy [9].

Since lexical chains were first proposed by Morris and Hirst [10], they have been used to address a variety of Information Retrieval (IR) and NLP applications, such as term weighting for IR tasks [15], malapropism detection [14], hypertext