Experiments on Statistical and Pattern-Based Biographical Summarization

Horacio Saggion and Robert Gaizauskas

Department of Computer Science,
University of Sheffield,
Sheffield - S1 4DP - United Kingdom
Tel: +44-114-222-1947
Fax: +44-114-222-1810
{saggion, robertg}@dcs.shef.ac.uk

Abstract. We describe experiments on content selection for producing biographical summaries from multiple documents. The method relies on a set of patterns to identify descriptive phrases, an available co-reference resolution algorithm, and a greedy, corpus-based sentence deletion procedure for document compression. We show that in an automatic evaluation of content using ROUGE, the proposed method obtains very good performance.

1 Introduction

Extracting relevant information from massive amounts of free text about people in order to construct profiles or biographies is a challenging problem not only because it is very difficult to elicitate in a precise way what type of information about a person is relevant for a biography, but also because even if some types of information were known to be relevant (e.g. birthdate or profession), there are many ways of expressing that information in natural language texts. As free text is by far the main repository of human knowledge, solutions to the problem of extracting information about people have many applications in areas of knowledge management and intelligence:

- In intelligence analysis activities: there is a need for access to personal information in order to create briefings for meetings; and for tracking activities of individuals in time and space;
- In journalism: there is a need to find relevant information for writing backgrounds or profiles for the main actors of a breaking news story;
- In publishing: the need is to update/create entries about famous people in Encyclopedias and dictionaries;
- In knowledge engineering: ontologies and other knowledge repositories need to be populated with instances such as persons and their attributes extracted from text.

C. Bento, A. Cardoso, and G. Dias (Eds.): EPIA 2005, LNAI 3808, pp. 611–621, 2005.

Recent natural language processing challenges such as the Document Understanding Conferences (DUC) and the Text Retrieval Conferences (TREC) Question Answering (QA) evaluations have focused on this particular problem and are creating useful language resources to study the problem and measure technical advances. For example, in task 5 in the recent DUC 2004 system participants had to create summaries from sets of documents answering the question "Who is X?", and from 2003 onwards, the TREC/QA evaluations have a specific task which consists of finding relevant information about a person in a massive text repository.

In this paper, we concentrate on the problem of creating a profile or biography for a particular person given a set of documents referring to that person as defined in the DUC 2004 evaluation. This is an instance of the more general summarization problem (see [1] for an overview of the field).

A multidocument summarization system takes as input a set of documents (or *cluster*) related by some "topic" and produces a summary of the whole set. If the cluster of documents refers to a given event instance (e.g. a particular earthquake), it is likely that sentences from different documents in the cluster will report the same information (e.g. damages, victims). Therefore, multi-document summarization algorithms can take advantage of the *redundancy* of information in order to measure relevance. However, when the collection of documents refers to a particular person other techniques seem to perform better than redundancy alone. For example, one of the best performing systems in the recent DUC 2004 evaluation (task 5) used syntactic criteria alone to select sentences in order to create a biographical summary [2]. Only two types of construction were used in that work: *appositive* and *copula* constructions both of which rely on syntactic analysis.

In the experiments to be reported in this paper, we focus on the problem of *extracting* the necessary information to create the person's profile; the problem of synthesis – the production of a coherent and cohesive biography – will not be discussed in this work. We take as a point of departure the ideas used in [2], however we follow a shallow pattern-based approach to the identification of profile information combined with a greedy search algorithm informed by corpus statistics. We will show that, unlike other methods, the proposed solution is ranked consistently high on the DUC 2004 Task 5 data. The rest of the paper is organized as follows: in the next section we describe the natural language processing tools and in Section 3 the process of content selection. Section 4 gives details of the process of document reduction. In Section 5 experiments using DUC 2004 data and results are presented. In Section 6 we report on past work and Section 7 closes with our conclusions and future work.

2 The System

We have developed a sentence selection mechanism which, given a target person and a cluster of documents referring to the target, extracts relevant content from the cluster and produces a summary such as the one presented in Figure 1.