ProGenIE: Biographical descriptions for Intelligence Analysis

Pablo A. Duboue, Kathleen R. McKeown, and Vasileios Hatzivassiloglou

Columbia University,
Dept. of Computer Science
New York, NY, 10025, USA
{pablo,kathy,vh}@cs.columbia.edu
http://www.cs.columbia.edu/nlp/

Abstract. Intelligence analysts face the need for immediate, up-to-date information about individuals of interest. While biographies can be written and stored in text databases, we argue that they can get quickly obsolete for living persons. We present here the architecture of ProGenIE, a biographical description generator currently under construction, focusing on the requirements of the task and its impact on potential users.

1 Introduction

Intelligence and law enforcement personnel face the need for immediate, up-to-date information about individuals of interest. Biographies or profiles can be used to present such information. Obviously, it is not feasible to have human writers produce biographies for each person at hand. Even if that were the case, such biographies would need to be later stored in textual databases, where they will lose track of most recent, and, arguably most important, activities of the person being described. As part of the joint Columbia University—University of Colorado Open Question Answering project (AQUAINT), we present here our proposed architecture for a biographical description generator from knowledge sources and information extracted from the Internet.

Our goal is to provide intelligence and law enforcement personnel with means to quickly and concisely communicate information about military and political personnel from foreign countries, and also terrorists and criminals. Working on different scenarios, different users of the system will require different presentations of available data about a given individual. For example, one analyst might want to see an overview of all data for a particular person, while another analyst may be looking for ties between a well-known terrorist and a particular country. We intend to fulfill these requirements via *on the fly* generation of such person descriptions.

This paper is organized as follows: we shortly describe the motivation and relevance of our system. In Section 3, we describe ProGenIE's three major components. Some final remarks conclude this paper.

2 Motivation and Relevance

Person descriptions has been addressed in the past by IR, summarization and NLG techniques. IR-based systems [1] will look for existing biographies in a large textual database such as the Internet. Summarization techniques [2] produce a new biography by integrating pieces of text from various textual sources. Natural language generation systems for biography generation [3] create text from structured information sources. Ours is a novel approach, that builds on the NLG tradition. We will combine a generator with an agent-based infrastructure expecting to ultimately mix textual (like existing biographies and news articles) as well as non-textual (like airline passengers lists and bank records) sources. ProGenIE will offer significant advantages, as pure knowledge sources will be able to be mixed directly with text sources and numeric databases. It diverges from the NLG tradition, as we will use examples from the domain to automatically construct content plans. Such plans will guide the generation of biographies on unseen people. Moreover, the output of the system will be able to be personalized; and by the fact that the system learns from examples, it will be able to be dynamically personalized.

3 System Description

Three components make up for our system: a knowledge component, a learning component (our research focus) and a generation component.

Learning Component. The key to greater flexibility in biography generation relies in a particular piece of the generation pipeline, the **Content Planner**. A content planner is responsible for the distribution of the information among the different paragraphs, bulleted lists, and other textual elements. Information-rich inputs require a thorough filtering, resulting in a small amount of the available data being conveyed in the output. The selection and structuring of the text, performed thus by the content planner, is responsible for our sought flexibilities.

Our research objectives focus on the automatic acquisition of **schemas**, data structures that guide the content planning process [4], by means of machine learning techniques. We employed an *aligned corpora* of input data and output text to induce schemas using stochastic search [5]. Such schemas are then used to generate biographies on new people, different from the one used to learn them. The final system can then be easily customized for new needs or scenarios by the final users, expanding the current work to possibilities unforeseen by us.

Knowledge Component. While the data employed to generate the biographies can be supplied by internal databases and networks such as Intelink or IAFIS [6], we plan to provide input to the generator by using information extraction agents on the Internet. Publicly available data can be of great use to mine information for well-known personalities and a test bed for the final system running on private intranets.

To represent the input to the generator, we chose a variation of **RDF**. This selection strives for generality, reuse and portability.

Generation Component. Seven modules in a pipeline will compose PROGENIE. These modules will include an Inference Module, a Content Planner, a Text Planner, a Referring Expression Generator, an Aggregation Module, a Lexical Chooser and a Surface Realizer. In this setting, the Content Planner module executes the learned schemas. We use a variation of the Lexical Chooser (that selects words for concepts) from the MAGIC generator and the FUF/SURGE unification based package for the Surface Realizer. Finally, the other modules will behave as follows: the Inference Module performs some limited world knowledge inferencing; the Text Planner splits a rhetorical tree into paragraphs; the Referring Expression Generator handles mostly pronominalization, although it can scan the input to generate descriptions like *his father*; and the Aggregation modules is responsible for mixing together clauses with similar structure, in order to avoid repetition.

4 Final Remarks

We have presented here a biography generation system with three components. A prototype of the learning component inferred plans for an earlier domain [5]. A new version of it, focusing in selecting appropriate pieces of content for a given biographical task, succeeded in halving the available data by two, while keeping the correct data for further verbalization [7]. The generation component has currently five operational modules, at different levels of completion. The remainder two modules are the Lexical Chooser, undergoing knowledge acquistion, and the Aggregation Module, on the design phase. PROGENIE solves an existing requirement for intelligence and law enforcement personnel. Its design has highly benefited from interviews with potential users; this fact is reflected on the architecture presented here. We plan to have an integrated biographies generator by the end of 2003, operating over publicly available Internet sources.

References

1. Müller, A., Kutschekmanesch, S.: Using abductive inference and dynamic indexing to retrieve multimedia sgml documents. In: Miro '95. (1995)
2. Schiffman, B., Mani, I., Conception, K.J.: Producing biographical summaries: Combining linguistic knowledge with corpus statistics. In: ACL-EACL 2001. (2001)
3. Teich, E., Bateman, J.A.: Towards an application of text generation in an integrated publication system. In: Proc. of 7th IWNLG. (1994)
4. McKeown, K.R.: Text Generation: Using Discourse Strategies and Focus Constraints to Generate Natural Language Text. Cambridge University Press (1985)
5. Duboue, P.A., McKeown, K.R.: Content planner construction via evolutionary algorithms and a corpus-based fitness function. In: INLG-2002. (2002)
6. U.S. Dept. of Justice, F.B.I.: Inauguration of the integrated automated fingerprint identification system (IAFIS). Press Release (1999)
7. Duboue, P.A., McKeown, K.R.: Statistical acquisition of nlg content selection rules. submitted (2003)