Chapter IV
Mining Profiles and Definitions with Natural Language Processing

Horacio Saggion
University of Sheffield, UK

ABSTRACT

Free text is a main repository of human knowledge; therefore, methods and techniques to access this unstructured source of knowledge are of paramount importance. In this chapter we describe natural language processing technology for the development of question answering and text summarisation systems. We focus on applications aiming at mining textual resources to extract knowledge for the automatic creation of definitions and person profiles.

INTRODUCTION

Extracting relevant information from massive amounts of free text about people, companies, organisations, locations, and common terms in order to create definitions or profiles is a very challenging problem not only because it is very difficult to elucidate in a precise way what type of information about these entities is relevant for a definition/profile, but also because even if some types of information were known to be relevant, there are many ways of expressing them in natural language texts. As free text is by far the main repository of human knowledge, solutions to the problem of extracting definitional information 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, broadcasting, and news reporting activities:** there is a need to find relevant information for writing back-
grounds for the main actors of a breaking news story, but also term definitions need to be provided to non-specialist audiences (e.g., *What is bird flu*?);

- **In publishing**: Encyclopaedias and dictionaries need to be updated with new information about people and other entities found in text repositories;

- **In knowledge engineering**: ontologies and other knowledge repositories need to be populated with instances such as persons and their attributes extracted from text, but also new terms together with their definitions need to be identified in texts in order to make informed decisions about their inclusion in these knowledge repositories;

- **In business intelligence**: information about companies and their key company officers is of great relevance for decision making processes such as whether or not to give credit to a company given the profile of a key company director.

Recent natural language processing challenges such as the document understanding conferences (DUC) ([http://www-nlpir.nist.gov/projects/duc/](http://www-nlpir.nist.gov/projects/duc/)) and the text retrieval conferences question answering track (TREC/QA) evaluations ([http://trec.nist.gov/data/qa.html](http://trec.nist.gov/data/qa.html)) 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 onward, the TREC/QA evaluations have a specific task which consists of finding relevant information about a person, an organisation, an event, or a common term in a massive text repository (e.g., *What is X?*).

In the Natural Language Processing Group at the University of Sheffield we have been working on these problems for many years, and we have developed effective tools to address them using the general architecture for text engineering (GATE). The main purpose of this chapter is to study the problem of mining textual sources in order to find definitions, profiles, and biographies. This chapter provides first an overview of generic techniques in natural language processing to then present two case studies of the use of natural language technology in DUC and TREC/QA.

## NATURAL LANGUAGE PROCESSING TOOLS

The General Architecture for Text Engineering (GATE) is a framework for the development and deployment of language processing technology in large scale (Cunningham, Maynard, Bontcheva, & Tablan, 2002). It provides three types of resources: language resources (LRs) which collectively refer to data; processing resources (PRs) which are used to refer to algorithms; and visualisation resources (VRs) which represent visualisation and editing components.

GATE can be used to process documents in different formats including plain text, HTML, XML, RTF, and SGML. When a document is loaded or opened in GATE, a document structure analyser is called upon which is in charge of creating a GATE document, an LR which will contain the text of the original document and one or more sets of annotations, one of which will contain the document markups (for example, html).

Annotations are generally updated by PRs during text analysis, but they can also be created during annotation editing in the GATE GUI (see Figure 1 for the GATE GUI). Each annotation belongs to an annotation set and has a type, a pair of offsets (the span of text one wants to annotate), and a set of features and values that are used to encode the information. Features (or attribute names) are strings, and values can be any Java object. Attributes and values can be specified in an annotation schema which facilitates validation and input during manual annotation. In Figure