The Twenty-One Demonstrator

Djoerd Hiemstra
University of Twente/CTIT
P.O. Box 217, 7500 AE, Enschede
The Netherlands, Tel: +31 53 4892335
hiemstra@cs.utwente.nl

Franciska de Jong
University of Twente/CTIT
P.O. Box 217, 7500 AE, Enschede
The Netherlands, Tel: +31 53 4894193
fdejong@cs.utwente.nl

Wessel Kraaij
Netherlands Organization for Applied Scientific Research (TNO)
Stieltjesweg 1, 2628 CK, Delft
The Netherlands, Tel: +31 15 2692259
kraaij@tpd.tno.nl

Abstract

Twenty-One is a project funded within the EU Telematics Applications Programme, sector Information Engineering. In Twenty-One non-profit environmental organisations, companies and research organisations work together to make documents on sustainable development and the environment more easily accessible. The most important deliverable of the project is the Twenty-One disclosure and retrieval system which produces an index on the multilingual multimedia document base. This index will be available via CD-ROM or WWW and accessible via a web-server. (`http://twentyone.tpd.tno.nl/`)

Keywords Full Text Retrieval; Cross-Language Retrieval; Domain Tuning; Multimedia Databases; Noun Phrase Indexing.

1 Introduction

Imagine yourself, sitting behind a PC that is connected to the Internet. You are looking for information on some environmental issue: acid rain, tropical woods, solar energy, recycling, or whatever. You have found some documents in your own language but they do not satisfy you. An exploring and curious person as you are, you would like to browse through documents in another language that is not your mother tongue, even if you don not understand a word of this other language. Suppose this other language is French. You want to look on the Internet whether there are French documents about the topic you are interested in. Imagine a computer system that allows you to type in a query in your own language, and get back documents, originally in French, but translated by the system in your language, so that you can read them! Current translation technoloy is unsufficient for this purpose. Still, the scenario described above is a realistic one. But how can it be done?

The answer is to simply accept a crippled syntax, because in information retrieval it is not the syntactic correctness of a text that is important but the relevance of the information to the users need. To judge whether a piece of information fits the users requests, there

is no need for a grammatical correctness. Moreover, if the information appears to be important, but is too ungrammatical or inunderstandable, the user can get it translated manually. In this way only those texts are manually translated that are actually used.

Translation might also seem a strange request in a virtual world called Internet that is dominated by the English language, nowadays generally accepted as the standard scientific and cultural language by the whole western world. But the real world is different. The majority of the people are still very resistant or incapable to cross the language borders. Moreover, a lot of very interesting information, especially in the environmental field, is simply not available in English. This information remains veiled for most people that do not speak the language in which the information is stated. Twenty-One aims at changing this situation rigorously without the mega-effort of translating every piece of information into every possible language. By making use of modern natural language technology, documents will be translated fully automatically into the users language, whatever this language may be. For the moment however, Twenty-One will be restricted to English, Dutch, French and German, with a tentative extension to Spanish.

2 Functionality

Twenty-One is much more than automatic translation. The project aims at multimedia objects, rather than just pieces of textual information. Twenty-One can be characterized by the following keywords:

Multimedia; The Twenty-One system aims at the disclosure of documents of different media types and / or data formats e.g. paper documents, WEB documents, word processor documents, text annotated images, audio or video material with textual annotations.

Document conversion; The system incorporates a component for the conversion of the various document formats into standard representation (SGML/HTML), including a tool for the conversion of paper documents into electronic format, on the basis of lay-out semantics analysis and optical character recognition.

Advanced disclosure techniques; The Twenty-One Multimedia document base will be disclosed using several advanced techniques like fuzzy matching, rule-based natural language processing for phrase indexing, relevance ranking and automatic hyperlinking.

Multilinguality; The Twenty-One database consists of documents in different languages, initially Dutch, English, French and German but extensions to other European languages are envisaged. Twenty-One supports Cross-Language Information Retrieval, that is users are able to retrieve documents in the four project languages by typing a query in their native language.

Domain-tuning: Sustainable Development; The aim of the project is to build a system that supports and improves dissemination of information about 'local agenda 21' initiatives. This requires a special effort in the acquisition of linguistic resources that are tuned to the language and vocabulary in this domain.

The project also provides a so-called Information Transaction Model that treats environmental multimedia information in such a way that both producers, providers and users of the information benefit from participating.

3 Background

In 1992 the UNCED Conference in Rio de Janeiro, visited by more than 170 countries all over the world, yielded a document called Agenda 21. This document outlines the basic principles and strategies for setting up sustainable development projects, and has become one of the most important documents for environmental organisations and local authorities all over the world. The conference had a follow-up in 1995 in Berlin. But two years earlier, in 1993, the European Commission started a huge Awareness Programme in Europe, to help local authorities, non- governmental organisations (NGOs) and citizens become aware of their potential role in sustainable development of their own environment. Basic issue within this program was (and is): how can we make one party make use of the experience of another party. For example: if a city in Holland has successfully carried out a project on vandalism, and the results might be applicable to a city in Italy, how and when does this information go from one place to another. Within the European Awareness Programme networks and protocols were developed to monitor the collection and distribution of local sustainable development projects and initiatives. But there is still virtually no technical infrastructure. Moreover the language borders and concurrent required translation of documents make it far too expensive to achieve the desired distribution. The Twenty-One project started in January 1996 as a European project sponsored by the Telematics Programme of the European Commission, Sector Information Engineering to fill in this gap. By automatic disclosure and translation of documents valuable information from any source about any subject in any language becomes within reach of even the smallest wallet.

4 The Twenty-One Demonstrator

By july 1997 the Twenty-One project will be halfway its duration. An intermediate demonstrator has been realized already. It contains most of the functionality mentioned above. The demonstrator allows users to look for documents in English and Dutch, to search for noun phrases, to search for similar documents (relevance feedback) and to look at bitmaps of the original (paper) versions of the documents. The demonstrator is available via the Twenty-One Web-page on: `http://twentyone.tpd.tno.nl/`

5 Partners in Twenty-One

Partners in Twenty-One are:

Industrial partners: Getronics (main contractor, The Netherlands), Highland Software Systems (Schotland) Rank Xerox (France),

Research Organisations and Universities: TNO-TPD (The Netherlands), DFKI (Germany), University of Tuebingen (Germany), University of Twente (The Netherlands)

Environmental organisations: Stichting MOOI (The Netherlands), Friends of the Earth (Belgium), VODO (France), Environ Trust (United Kingdom), Climate Alliance (Germany)

For more information please contact: Dr. W.G. ter Stal, Project Coordinator Twenty-One, Getronics Software, PO Box 22678, Amsterdam, the Netherlands, phone: +31-20-4306126, telefax: +31-20-4306030