Semi-automatic indexing system for technical abstracts

Fact Sheet

Project Information

SISTA

Grant agreement ID: LRE61040

Project closed

Start date
1 January 1993

End date
1 March 1995

Funded under
Specific programme of research and technological development (EEC) in the field of telematic systems in areas of general interest - Linguistic research and engineering -, 1990-1994

Total cost
No data

EU contribution
No data

Coordinated by
Brian Training Ltd
United Kingdom

Objective

The SISTA project addresses the area of Document Abstracting and Indexing. SISTA is a 2 year project to develop an NLP-based framework to assist in automatic indexing of technical abstracts written in English.

The project's major deliverable will be a PC-implemented prototype that automatically reads a technical abstract, isolates and priorities its main index terms and marks the location of the terms within the abstract. An interactive tool then allows Elsevier's indexers to confirm or edit the terms and their locations. The overall

indexing system will therefore the semi-automatic and will represent a realistic compromise between productivity and accuracy.

Drawing on the consortium's expertise in abstract publishing, NLP research and advanced PC software development, SISTA's central commercial objective will be to research, design and build a suite of prototype PC-based tools to index, semi-automatically, technical abstracts written in English.

The commercial purpose of the tools is to:

improve the productivity and consistency of the indexing process;
improve the quality of index users' facilities;
contribute to the equality of access ideal for databases, especially text databases in machine-readable form, covering arbitrary subject matter in unrestricted discourse;
provide an NLP-based framework for more diverse applications such as on-line documentation and technology-based training.

Although SISTA focuses on English as its application language, a major concern will be to develop a solution that will allow the methodology to transfer directly to other languages.

SISTA's indexing problem is provided by Elsevier Science Publishers BV who believe that immediate improvements in productivity and consistency can be achieved by developing current NLP technology in a carefully targeted way. Being user-led, the Consortium has set itself commercially significant but technologically feasible goals whose impact can be assessed objectively.

SISTA's NLP technology represents a novel combination of classical symbolic processing and statistical analysis. Firstly, each pre-indexed abstract undergoes a surface-oriented parse which reveals basic sentence structure. Next, potentially diagnostic constituents of the sentence structure are isolated. Then, by comparing these diagnostic constituents with the correct index entries, across the corpus of pre-indexed abstracts, a statistical model is developed. Lastly, the model's robustness is tuned by a global text matching algorithm. The final system will therefore take an abstract as input and return a prioritised list of index terms. The location of the terms within the abstract is then marked, using the ISO standard SGML notation, according to linguistic data provided by the parser.

The technological challenge is to provide, for each abstract, a prioritised list of index terms, drawn from a thesaurus. The strategy is to use a large corpus of pre-indexed abstracts to train a statistical model. This model will then tested against a second set of pre-indexed material. The refined model will then be give a prototype PC-implementation which will be evaluated on-site at Elsevier.

Clearly SISTA's technological strategy presupposes experience in both symbolic processing and stochastic methods, not only at the level of academic research, but also in building commercial implementations. These key skills are represented strongly both at the research and the technology transfer level.

Although SISTA focuses on a specific well-defined commercial problem, the proposed solution has widespread generic application beyond publishing to such diverse sectors as technology-based training and on-line documentation. The beneficiaries of a semi-automatic indexing tool are index developers and index users. Improvements in productivity will bring direct resource savings to index developers. In addition, it is one of SISTA's aims to evaluate any improvements in indexing consistency and quality. Improved consistency will bring direct savings by reduced user time and indirect benefits of increased ease of use to end users.

The most direct qualitative result will be the development of improved text retrieval methods by importing computational linguistics techniques into current information retrieval technology. This process will also inform and benefit current computational linguistics research, since it will allow the validation of tools in real-life text processing applications.

It is hoped that SISTA will contribute to the process of change in computational linguistics by demonstrating that transfer of NLP technology can yield linguistically well-motivated yet robust systems. In particular, SISTA's twinning of NLP technology with the text representation formalism of SGML may stimulate collaboration between these historically separate research communities. It is hope that this stimulus will be felt by researchers working on European languages other than English.

Although for the purposes of the project SISTA's subject domain is technical abstracts, the project's techniques will apply directly to abstracts dealing with other topics. In addition, SISTA's results will prove useful to other quite different applications and growing markets such as on-line documentation and technology-based training, both in English and other languages.

Fields of science (EuroSciVoc)

humanities > languages and literature > **linguistics**

natural sciences > computer and information sciences > **databases**

natural sciences > computer and information sciences > software > **software development**

social sciences > economics and business > economics > production economics > **productivity**

natural sciences > mathematics > applied mathematics > **statistics and probability**

Programme(s)

FP3-LRE - Specific programme of research and technological development (EEC) in the field of telematic systems in areas of general interest - Linguistic research and engineering -, 1990-1994

Topic(s)

Data not available

Call for proposal

Data not available

Funding Scheme

Data not available

Coordinator

Brian Training Ltd

EU contribution

No data

Total cost

No data

Address

Jeffreys Building St John's Innovation Centre Cowley Road
CB4 4WS Cambridge
United Kingdom

Participants (1)

AIC Ltd

Ireland

EU contribution

No data

Address

Total cost

No data

Last update: 7 July 1993

Permalink: https://cordis.europa.eu/project/id/LRE61040