A System of Automatic Textual Abridgement

Antoinette Renouf

Research and Development Unit for English Studies
University of Liverpool
19 Abercromby Square
Liverpool L19 OPJ

tel: *+44 0151 794 2286*
fax: *+44 0151 794 2298*
email: *ajrenouf@liverpool.ac.uk*

Abstract

We have developed a system which automatically produces abstracts, or abridgements, from electronic texts. It can abridge any text with normal features, ranging from newspaper articles to complete books. The resultant abridgement consists of a set of core information-bearing sentences, the unique feature of which is that they together also form a coherent and readable mini-text. The system is fast and efficient. Various parameters may be set by the user, including length. Among additional facilities to improve performance is one designed to resolve pronominal or other ambiguity. The system works on other languages, and has applications in IT and beyond.

Keywords

linguistics, NLP, software, text indexing, automatic abstracting, abridgement, information retrieval, translation aid, textual database interface, multilingual interface

Domains

text indexing, automatic abstracting, information retrieval, writing aids, translation aids, multilingual interfaces

A System of Automatic Textual Abridgement

Antoinette Renouf

Extended Abstract

Description of System

Over a period of several years, we have been developing and refining a system which automatically produces abstracts, or abridgements, from articles or books. We use the term 'abridgement' because the product is extracted from the original text rather than re-written. In collaboration with Michael Hoey, the author of the original idea, we have created a system which is now very fast and efficient.

Any kind of text can be abridged, provided that it has the normal features of text, in terms of linked sentences and an identifiable theme. Listings, for example, would not meet the requirement. Typical candidates are newspaper articles, but longer documents and entire books can also be processed - a book of 300 pages in under a minute. The length of abridgement may be varied, from one sentence upwards; the optimal length depends on the information content of the original. By default, it will typically be around one-fifth of the original. Other parameters which may be varied by the user include stopword and lemmatisation facilities.

The abridgement consists of a set of core information-bearing sentences from the text, identified as core by their lexical content and their interrelationship with the content of other sentences. They are therefore lexically rich, though not necessarily the longest sentences in the original. A unique feature of the system is that the core sentences are by definition cohesive with each other, and so form a mini-text which, in addition to summarising the original, is a coherent entity in itself.

As expected, there are some potential inhibitors to the readability of the abridgements. These include unresolved pronouns and ellipsis. The degree of problem which they pose varies according to text type, but most ambiguity occasioned by these elements is usually removed by a 'windowing' facility, which prints the sentences surrounding the ambiguous sentence in the original text wherever required. A more minor irritant is the occasional discourse signal, such as *however*, which may have become redundant in the new context of the abridgement. These elements are usually removable.

The system works with other languages, and has been tested on French, Dutch, Portuguese, Spanish and Greek texts. With appropriate modification, it will therefore have direct applicability in the multi-lingual domain.

The software is written entirely in ANSI C code. It runs on Unix, but the code is portable across different platforms, and also runs on a PC. The current version of the system is in full working order, and we are in the process of packaging it for market. As the system stands, the abridgement can function either as a final artefact, an intermediate product, or as something between the two. It has obvious applications in the fields of text retrieval and translation.

Some IT Applications

Textual database search, even with the various enhancements to basic keyword search now available, yields a proportion of irrelevant texts which take time to eliminate. A more efficient approach will be for the database user to be presented with an abridgement as a first-stage response to an enquiry, to allow a decision to be made as to whether or not to go ahead with retrieval of the full text.

Furthermore, it is entirely possible with this system to reduce an entire database to a series of abridgements, and thus to reduce search time to a fraction of current norms. This will be

a minimally information-losing exercise, since the abridgements retain the key lexis of the original text. The facility to retrieve full text if required could, however, be retained.

Some Applications in Translation

As a translation aid, an abridgement can give the translator or indexer an overview of the propositional content and key lexis of the original text. It can also be used for training purposes. In addition, the mechanisms which create the abridgements can be used to verify translations, in that key sentences in the source text should be reflected as key sentences in the target text.

Abridgements can also serve a useful purpose in reducing unnecessary volume in the many multi-lingual, parallel text archives currently being created and stored by international bodies, notably within the EU. As new alliances are formed, versions of the same document have to be translated into an increasing number of languages. Instead of translating and holding full translations of the source text, it may be sufficient for the translators to produce translations only of the abridged form of the original.

Further Development

There are inevitably further refinements which could be made. Tailoring the software to the individual needs of users is an obvious one. As a research unit, we are keen to carry out further enhancements, particularly those involving fundamental research effort. We have a clear idea of developments which interest us, and these range from the enhancement of the abridgement system per se to its adaptation as a tool in various IT applications.

We are interested in securing European industrial partners to work with us as research partners. The first stage of collaboration is likely to involve the submission of a proposal to carry out research with funding under the language engineering sector of the Telematics Applications Programme. Direct industrial sponsorship might be another course to consider. Any research which we undertake collaboratively will have the current abridgement software as a starting point. We are also interested in the ideas of collaborators and prepared to design a research programme that would be of joint benefit.