Automated, rule-based extraction of terms and definitions

Glossaries are often incomplete, time-consuming, and difficult to build. Many times when writing a technical paper, authors have to find the correct balance between assuming that the reader will understand a term and explicitly providing the definitions of the terms used in papers. This method allows authors to focus on communicating the essential ideas without being bogged down in considering which terms should be defined and which ones need not be. This technology automatically finds pertinent terms in text and then locates the definitions for these terms. After finding the terms and definitions, a glossary is automatically built. This technology can be used to make a glossary for any input text. All the important terms can be defined automatically, wasting none of the author’s time and allowing readers who may not be as familiar with the terminology to understand the terms and definitions.

Identifying definitions and their technical terms from online text

This technology can automatically identify and extract medical terms along with their definitions and modifiers from text. It is aimed primarily towards consumer-oriented medical articles but has use in any setting where terms are necessarily defined. For example nearly 60% of definitions are introduced or delimited by markers such as “–” or “()”. Accordingly, this technology deploys a two-stage system that contains a rule-based pattern extractor using text markers and a natural language parser. The Unified Medical Language System (UMLS) was used for domain knowledge. This automatic technology obviates the need for the author of a paper to do the added work of building a glossary. It is able to handle a wide range of inputs from the user.

This software was tested using a MEDLINEplus article. The results were evaluated using one method to reflect the quality of identifying worth terms and another method to reflect the quality of paring terms with their correct definitions, performing highly in both situations.

Lead Inventor:

Judith L. Klavans, Ph.D.
Applications:
- Creation and/or enhancement of online terminological resources
- Summarization and text categorization according to level of expertise
- Present information to patients in a language that is easier to understand
- Medical research articles
- Consumer oriented medical articles
- Highly technical research articles
- Decreases the required reading level of a text without sacrificing detail for those who want it

Advantages:
- Saves time building glossaries
- Makes articles more readable and accessible
- Compatible with any reading level

Patent Information:
Patent Pending (WO/2003/027894)
Tech Ventures Reference: IR M01-030

Related Publications:

Inventors
Kathleen McKeown