Image Retrieval in Controlled English

Tobias Kuhn and Michael Krauthammer

The Yale Image Finder (YIF) project aims at improving biomedical image and document retrieval by developing advanced image parsing and indexing strategies. To this end, we have deployed a YIF search engine, which allows for keyword searches against indexed Pubmed Central open access images. Authors often follow well-accepted layouts when depicting experiment results as gels, graphs or plots, and use image text in an equally structured fashion for labeling different image elements. Image text placement often conveys higher-level semantics, such as the names of proteins being studied under different experimental conditions. We are currently exploring innovative ways for allowing YIF users to access such structured image text content. Here, we propose the use of a controlled language interface that guides users in composing natural language queries ("Find an image where X is measured under the condition Y") that are be subsequently mapped to indexed image text content.

Our approach is based on controlled natural language, i.e. a restricted subset of English with a precise and unambiguous mapping to logic. We present a prototype called Rice (Retrieving Images through Controlled English) that is based on an interface we developed for a different domain (annotated text corpora) and adopted for image mining. Users can write seemingly natural queries like "Find an image that is a Western blot and where 'p38' is compared to 'MKK3'" which is subsequently translated into a logical representation like "western-blot & compared(p38,MKK3)". Such logical representations can then be matched with the formal model that we extract from images found in biomedical papers.

One serious problem with controlled natural language is that it is very easy to read and understand but hard to write. Our prototype solves this problem by providing a predictive editor, with which users construct syntactically correct sentences in an iterative and guided way. For any partial sentence, the predictive editor of Rice shows the possible continuations in the form of different menu boxes. In this way, users do not need to know about the restrictions of our language beforehand. Previous evaluation has shown that this editor is very easy to use after very little or no training. Typical users of search engines are not familiar with logic notations and rarely have the time to learn one. Existing query interfaces are either very simple (i.e. keyword-based) or too complex to be usable without training. With Rice, complex queries can be written in a natural and intuitive way. The interface should be immediately accessible to researchers interested in the results represented in images of the biomedical literature. Rice supports queries with directed relationships "... where A is measured under the condition B", resulting in the retrieval of highly specific image sets. In contrast, keyword searches cannot build such refined query representations, and cannot easily tell apart a related query "... where B is measured under condition A". Our prototype is still incomplete, but we believe that it nicely demonstrates the potential of our approach, and the positive results of previous work make us confident of its practicality.