

Selection of Passages for Information Reduction *

Jody J. Daniels
Department of Computer Science
University of Massachusetts
Amherst, MA 01003 USA
Email: daniels@cs.umass.edu

There currently exists a bottleneck in extracting information from pre-existing texts to generate a symbolic representation of the text that can be used by a case-based reasoning (CBR) system. Symbolic case representations are used in legal and medical domains among others. Finding similar cases in the legal domain is crucial because of the importance precedents play when arguing a case. Further, by examining the features and decisions of previous cases, an advocate or judge can decide how to handle a current problem. In the medical domain, remembering or finding cases similar to the current patient's may be key to making a correct diagnosis: they may provide insight as to how an illness should be treated or which treatments may prove to be the most effective.

This thesis demonstrates methods of locating, automatically and quickly, those textual passages that relate to predefined important features contained in previously unseen texts. The important features are those defined for use by a CBR system as slots and fillers and constitute the frame-based representation of a text or case. Broadly, we use a set of textual "annotations" associated with each slot to generate an information retrieval (IR) query. Each query is aimed at locating the set of passages most likely to contain information about the slot under consideration.

Currently, a user must read through many pages of text in order to find fillers for all the slots in a case-frame. This is a huge manual undertaking, particularly when there are fifty or more texts. Unfortunately, full-text understanding is not yet feasible as an alternative and information extract techniques themselves rely on large numbers of training texts with manually encoded answer keys. By locating and presenting relevant passages to the user, we will have significantly reduced the time and effort expenditure. Alternatively, we could save an automated information extraction system from processing an entire text by focusing the system on those portions of the text most likely to contain the desired information.

This work integrates a case-based reasoner with an IR engine to reduce the information bottleneck. SPIRE [Selection of Passages for Information REduction] works as follows: the CBR system evaluates its case-base relative to a current problem situation. It passes along to the IR engine the identifiers of the documents that describe fact situations the most similar to the current problem. The IR engine treats these documents as though they were hand-marked as relevant and uses them to generate a query against a larger corpus of texts (Daniels and Rissland 1995).

After retrieving additional relevant texts, we might wish to add them to the CBR system's knowledge base. However, the documents must be converted from their original text into a frame-based representation, a time-consuming and error-prone activity. To assist the knowledge engineer, we save a set of "annotations", which we derive when creating the original case-base. An annotation contains the words and phrases that describe the value of a particular slot filler and the annotation is associated with its respective slot. An annotation may be a segment of a sentence, an entire sentence, or several sentences. For example, for the slot that contains the value of someone's *monthly-income*, sample annotations from SPIRE's case-base are: "net disposable monthly income for 1979 averaged $1,624.82" and "His current gross income is $24,000 per year."

SPIRE passes the case-base of annotations for each slot to the IR system. Using these annotations, the IR component generates a new query aimed at retrieving small relevant passages from the documents just retrieved. By combining into a query those descriptive terms and phrases used to identify the slot fillers within the current case-base, we can locate relevant passages within novel texts.

By retrieving passages for display to the user, we have winnowed a text down to several sets of sentences. This process is repeated for each slot in the case-based reasoner's representation of the problem. By locating and displaying these important passages to a user, we have reduced reading an entire document to examining several sets of sentences, resulting in a tremendous savings in time and effort.

* This research was supported by NSF Grant no. EEC-9209623, State/Industry/University Cooperative Research on Intelligent Information Retrieval, Digital Equipment Corporation and the National Center for Automated Information Research.

References

Jody J. Daniels and Edwina L. Rissland. A Case-Based Approach to Intelligent Information Retrieval. In *Proceedings of the 18th Annual International ACM/SIGIR Conference*, pages 238–245, Seattle, WA, July 1995. ACM.