Text pattern visualization for analysis of biology full text and captions

Andrea Grimes and Robert Futrelle

Northeastern University, College of Computer & Information Science

{agrimes, futrelle}@ccs.neu.edu

Overview

We have developed visualization software to aid in the discovery of language patterns in biology papers. These patterns can help to clarify how biologists package information linguistically. This is important for building efficient information retrieval systems and to reach the end goal of answering users' questions.

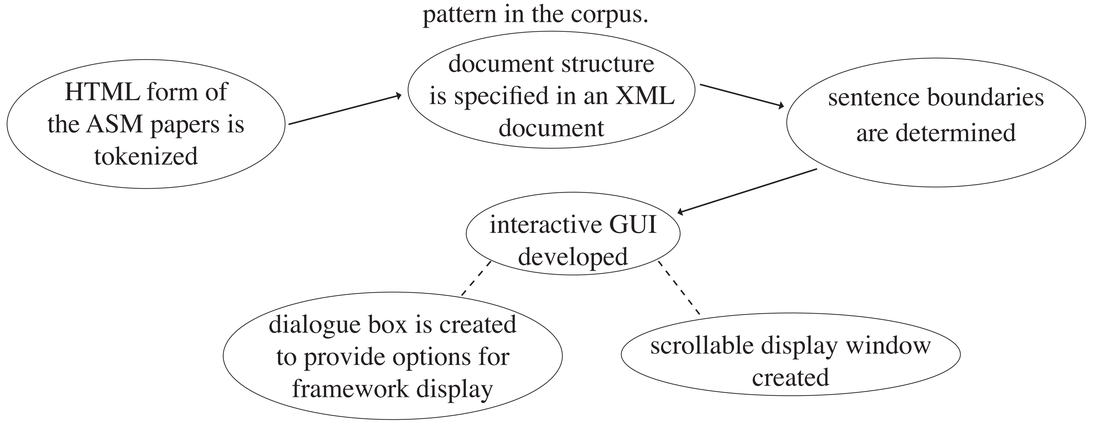
- Corpus Size: 250M words from 50K papers from ASM
- Studies thus far: 5M words from 1K papers
- Full text (paper body and captions) is analyzed: not just abstracts

Frameworks = Language Patterns

- the way in which biologists package information
- containers: consist of anchors (high frequency, low information) and endograms/exograms (low frequency, high information)
- used to extract information from text
- characterized by the set of informational elements that they induce

Framework Viewer (FV) Implementation

The FV is an interactive software system which allows the user to search for instances of a "Framework" pattern in the corpus.



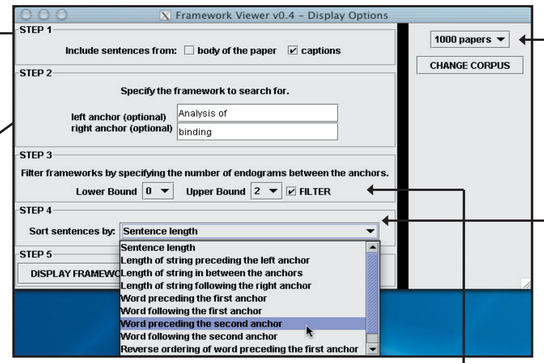
Framework Analysis

We view text as data rather than as logical structures. All instances of a Framework in the corpus are considered collectively. The set of highly informational words induced by the Frameworks is of primary interest. These classes are more informative than traditional part-of-speech classes. For instance, in our clausal framework analysis, we look at specific clausal categories such as descriptions of causality relations. The classes of words induced by causality-oriented Frameworks are compared for similarity and allow us to produce a set of words which function in causality Frameworks.

Framework Viewer (FV) - Display Options

The FV is able to search the body of the paper, the captions, or both.

Either 1 or 2 anchors may be specified. Each anchor can be any number of words and symbols. "<e>" may be typed to request that any "entity" be returned, e.g., "BRCA1" or "Hsp72".



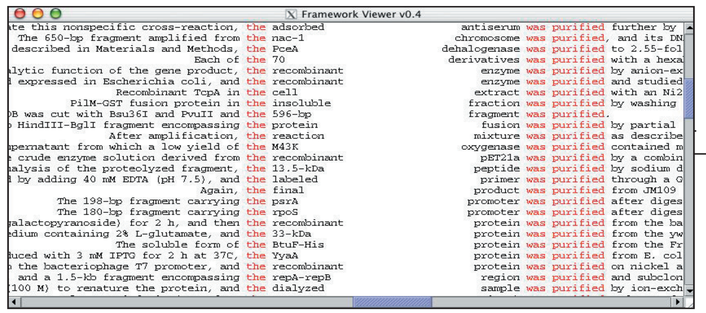
The choices of corpus size in the prototype FV are 1, 10, 100, or 1000 papers.

Various sorting options are available. A stable sorting algorithm is used so that multiple sorts may be done to customize the order in which the sentences are displayed.

A filter can be applied to the frameworks found so that only instances with the specified number of words between the anchors are returned.

Clausal Analysis Using the Framework Viewer (FV)

(studies done on a corpus of 1000 papers)



The display of the FV for the framework with vertically-aligned anchors "the" and "was purified". The instances were sorted first by the word preceding the right anchor and then by length of the string between the anchors. This framework is an example of a methodology clause and induces a set of words where the relation among the elements is that they were each purified.

The display of the FV for the framework "were treated with" in which the instances are sorted based on the sentence length. Only one anchor is specified here. The words which comprise the anchor are all of low information content, whereas the word(s) immediately following the anchor have high information content.