FlyBase Curation using Natural Language Processing

Nikiforos Karamanis*,†, Ruth Seal‡, Ian Lewin*, Andreas Vlachos*, Caroline Gasperin† and Ted Briscoe∗

*Computer Laboratory and ‡Department of Genetics

There has been much recent research on applying Natural Language Processing (NLP) techniques to biomedical text as a potential aid to database curation. However, developing integrated systems which address the curators’ real-world needs has been studied less rigorously.

As part of our work in the FlySlip project, we have been following a user-centered approach to develop PaperBrowser, a system which presents FlyBase curators with the output of several NLP processes on a paper-by-paper basis. PaperBrowser has been integrated into the curation workflow to help curators work more efficiently.

A user-centered approach to system development

PaperBrowser has been developed under an approach in which the curator is actively involved in every phase of software development. Initially, developers observed curators at work and following discussions, the following design requirements were identified for PaperBrowser:

• Provide easy access to full text.
• Enable curators to navigate the text quickly and efficiently.
• Direct their attention to relevant words in the text (such as gene names and related phrases).
• Allow them to maintain the initiative and follow their preferred curation style.

PaperBrowser

PaperBrowser presents the curator with an enhanced display of the text in which words automatically recognised as gene names are highlighted in a colored font. It is equipped with two navigation panes: PaperView, which is organised in terms of document structure and EntitiesView, which shows the relations between noun phrases that are relevant to curation. The nodes in the navigation panes are hyperlinked to the text to allow the curators to quickly and efficiently navigate through the paper in their search for curatable information.

PaperView lists automatically recognised gene names such as “dpp” in the order in which they appear in each section.

EntitiesView groups together noun phrases automatically recognised as referring to the same gene or to a biologically related entity such as “the zen cDNA”.

Evaluation

In our first evaluation study,² we compared the time it took the curator to curate a research paper using PaperBrowser to the time taken using a hard copy of a paper.

• The study revealed a trend for curation to be completed more quickly (by about 20%) when the curator was interacting with the paper via PaperBrowser.

In our most recent study,³ an evaluation task was devised in which two curators highlighted text that contains curatable information using PaperBrowser’s navigational aids. Navigating the text using the "Find" function served as the control condition.

• The study reveals that it takes about 58% fewer user actions to spot curatable text when PaperBrowser’s navigators are used.

After this study, one curator said:

“The highlighting, both for PaperView and EntitiesView, was very useful, and I found myself trusting it and using it to direct my curation.”

To date, over 75 papers have been curated with the help of PaperBrowser without aborting it (100% success rate).

Conclusion

State-of-the-art performance in tasks such as NER and Anaphora Resolution can be combined with the navigational and highlighting functionalities of PaperBrowser to support curation successfully.

Publications


The FlySlip project was funded by the UK Biotechnology and Biological Sciences Research Council (grant No 38688).

Webpage: www.cam.cl.cam.ac.uk/rowiki/NaturalLanguage/FlySlip