Imitating manual curation of text-mined facts

Raul Rodriguez-Esteban, Ivan Iossifov, and Andrey Rzhetsky

Center for Computational Biology and Bioinformatics,

Columbia University, New York

Introduction

Current automated approaches for extracting biologically important facts from scientific articles are imperfect. In order to emulate the human experts evaluating the quality of the automatically extracted facts, we have developed an artificial intelligence program ("a robotic curator") that closely approaches human experts. We illustrate our analysis by visualizing the predicted accuracy of the text-mined relations involving cocaine (see Fig. 1).

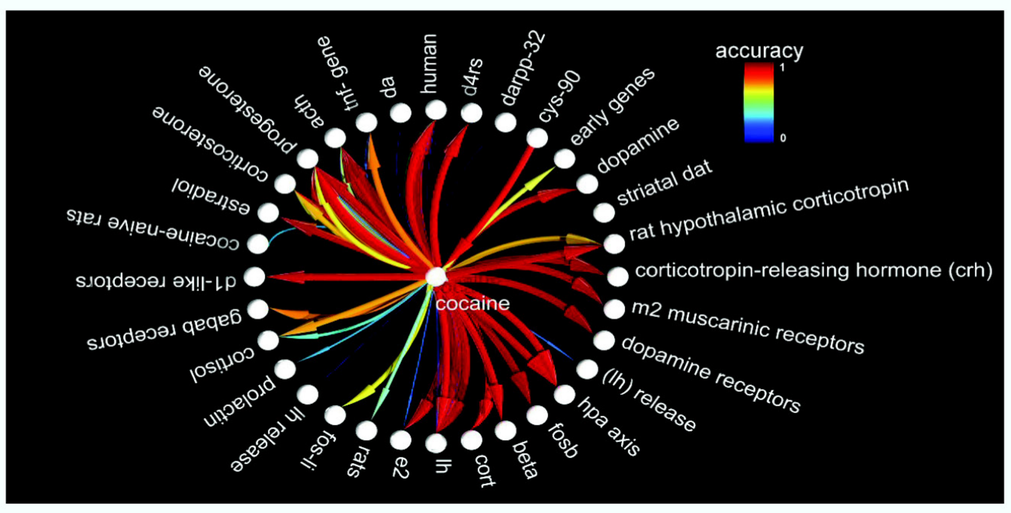


Fig. 1. Cocaine: predicted accuracy of individual text-mined facts involving semantic relation *stimulate*. The predicted accuracy of individual statements is indicated both in color and in width of the corresponding arc. Note that, for example, the relation between *cocaine* and *progesterone* was derived from multiple sentences.

Materials and methods

Our approach followed the path of supervised machine-learning. First, we generated a large training set of facts that were originally gathered by our information-extraction system [1], and then manually labeled as *correct* or *incorrect* by a team of human curators (about 100,000 annotations). Second, we used a battery of machine-learning tools to imitate computationally the work of the human evaluators, choosing a set of descriptive features (see Fig. 2). Third, we split the training set into ten parts, so that we could evaluate the significance of performance differences among the several competing machine-learning approaches.

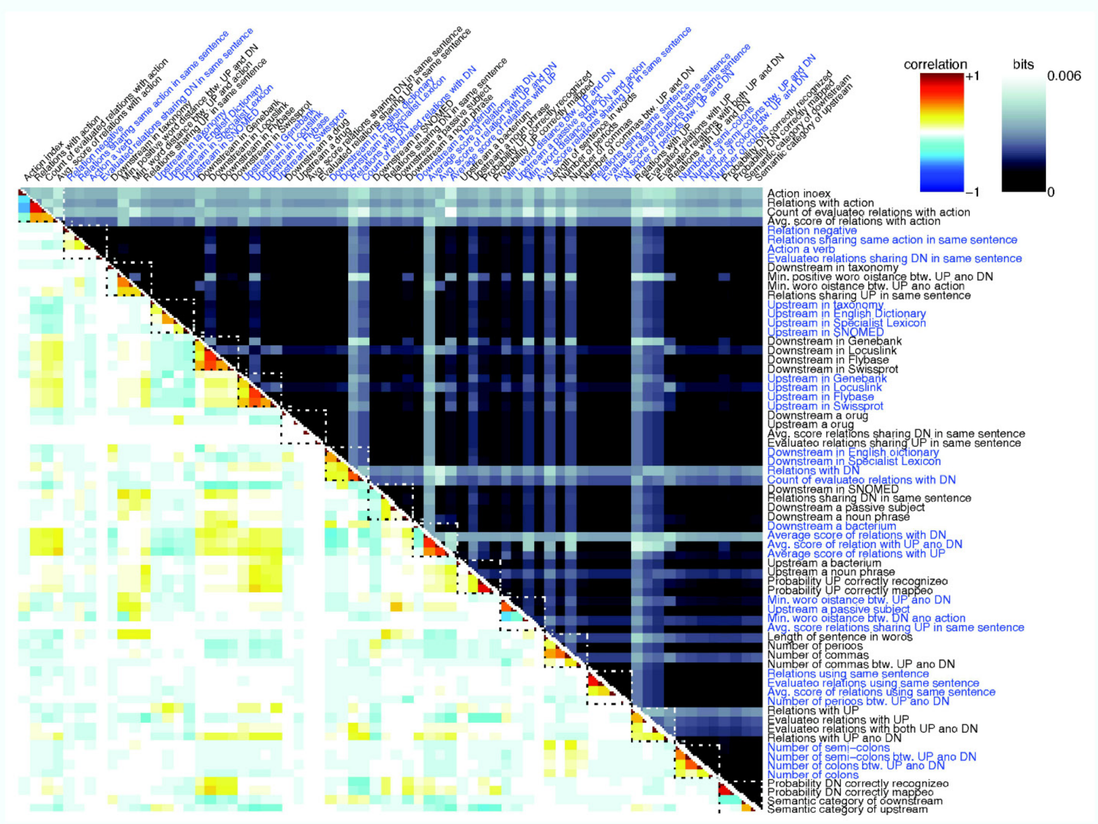
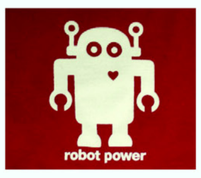


Fig. 2. Comparison of a correlation matrix for the features (colored half of the matrix) and a matrix of mutual information between all feature pairs and the statement class *correct* or *incorrect*. The plot indicates that a significant amount of information critical for classification is encoded in pairs of weakly correlated features. The white dotted lines outline clusters of features.



Results

Even the simplest Naïve Bayes method had an average ROC score of 0.84, which more sophisticated approaches surpassed to reach almost 0.95 (see Fig. 3-5).

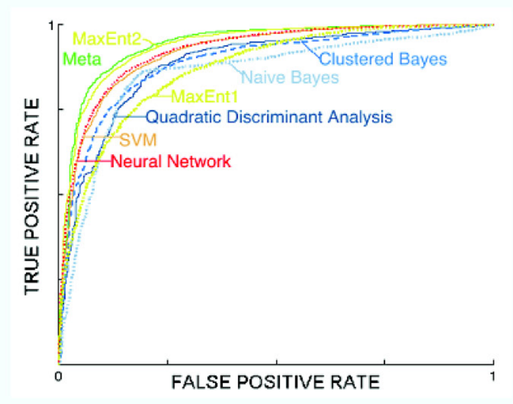


Fig. 3. Receiver-operating characteristic (ROC) curves for the classification methods that we used.

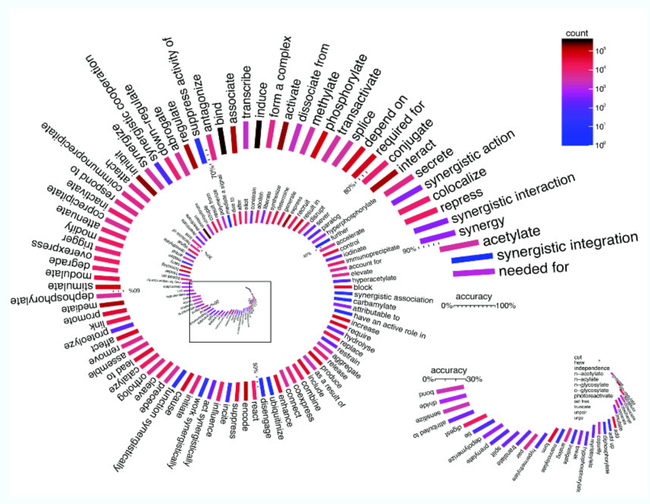


Fig. 4. Accuracy of the non-curated relations in the GeneWays 6.0 database. The plot compactly represents both the per-relation *accuracy* of the extraction process (indicated with the length of the corresponding bar) and their *abundance* (represented by the bar color).

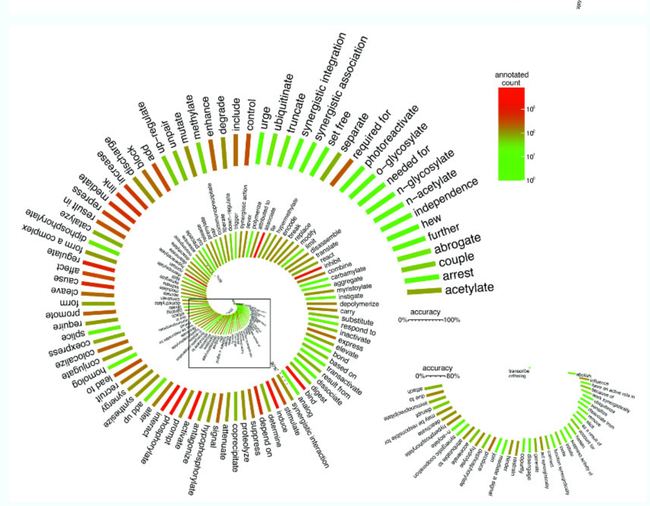


Fig. 5. Accuracy and abundance of the extracted and *automatically curated* relations. This plot represents both the per-relation accuracy after both information extraction and automated curation were done. Accuracy is indicated with the length of the relation-specific bars, while the *abundance* of the corresponding relations in the *manually curated* data set is represented by color.

Discussion

It is a matter of both academic curiosity and of practical importance to know how the performance of our artificial intelligence curator compares to that of humans. If we define the *correct* answer as a majority-vote of the human evaluators, the average accuracy of MaxEnt 2 is slightly lower than, but statistically indistinguishable from humans. If, however, in the spirit of Turing's test of machine intelligence [2], we treat the MaxEnt 2 algorithm on an equal footing with the human evaluators, MaxEnt 2 always performs slightly more accurately than one of the human evaluators.

Conclusion

Machine-learning curation of text-mined facts can reach the classification quality of human curators.

Literature cited

[1] Rzhetsky A, et al. (2004) GeneWays: a system for extracting, analyzing, visualizing, and integrating molecular pathway data. J Biomed Inform 37:43-53.

[2] Turing A (1950) Computing machinery and intelligence. Mind 59:433–560.

Acknowledgements

We are grateful to Mr. Marc Hadfield and Ms. Mitzi Morris for programming assistance. This study was supported by grants to A.R. from N.I.H. (GM61372), N.S.F. (supplement to EIA-0121687), the Cure Autism Now Foundation, and D.A.R.P.A. (FA8750-04-2-0123).

Further Information

This poster is based on a manuscript accepted for publication at the journal PLoS Computational Biology.

E-mail: raul@ee.columbia.edu . Homepage: http://www.ee.columbia.edu/~raul/

PDF version of the poster: http://www.ee.columbia.edu/~raul/IMC.pdf