Poster K-26

Imitating manual curation of text-mined facts in biomedicine

Authors:
Raul Rodriguez-Esteban (*Center for Computational Biology and Bioinformatics, Columbia University*)
Ivan Iossifov (*Center for Computational Biology and Bioinformatics, Columbia University*)
Andrey Rzhetsky (*Center for Computational Biology and Bioinformatics, Columbia University*)

Short Abstract: In biomedical applications, which rely on use of text-mined data, it is critical to assess the extraction quality of individual facts. Using a large set of almost 100,000 manually produced evaluations, we implemented algorithms that mimic human evaluation of facts provided by an automated information-extraction system.

Long Abstract:
Text-mining algorithms make mistakes in extracting facts from the natural-language texts.
In biomedical applications, which rely on use of text-mined data, it is critical to assess the extraction quality (the probability that the message is correctly extracted) of individual facts. Using a large set of almost 100,000 manually produced evaluations (most facts were independently reviewed more than once producing independent evaluations), we implemented and tested a collection of algorithms that mimic human evaluation of facts provided by an automated information-extraction system. The algorithms that were used include several Bayesian classifiers, SVMs, Neural Networks and Maximum Entropy methods. The performance of our best automated classifiers, a second-order Maximum Entropy classifier, closely approached that of our human evaluators (ROC score close to 0.95). Were we to use a larger number of human experts to evaluate any given sentence, we could implement an artificial-intelligence curator that would perform the classification job at least as accurately as an average individual human evaluator.
Hence we present a system that automatically curates the interactions that are extracted from the biomedical literature. This system is useful for enhancing the quality of information gathered by text-mining techniques. We illustrate our analysis by visualizing the predicted accuracy of the text-mined relations involving cocaine. We illustrate our analysis by visualizing the predicted accuracy of the text-mined relations involving cocaine.