Semi-Automated Text Classification

Giacomo Berardi
Istituto di Scienza e Tecnologia dell'Informazione
Consiglio Nazionale delle Ricerche
Pisa, Italy
giacomo.berardi@isti.cnr.it

Abstract

There is currently a high demand for information systems that automatically analyze textual data, since many organizations, both private and public, need to process large amounts of such data as part of their daily routine, an activity that cannot be performed by means of human work only. One of the answers to this need is *text classification* (TC), the task of automatically labelling textual documents from a domain $\mathcal{D}$ with thematic categories from a predefined set $\mathcal{C}$. Modern text classification systems have reached high efficiency standards, but cannot always guarantee the labelling accuracy that applications demand. When the level of accuracy that can be obtained is insufficient, one may revert to processes in which classification is performed via a combination of automated activity and human effort.

One such process is *semi-automated text classification* (SATC), which we define as the task of ranking a set $\mathcal{D}$ of automatically labelled textual documents in such a way that, if a human annotator validates (i.e., inspects and corrects where appropriate) the documents in a top-ranked portion of $\mathcal{D}$ with the goal of increasing the overall labelling accuracy of $\mathcal{D}$, the expected such increase is maximized. An obvious strategy is to rank $\mathcal{D}$ so that the documents that the classifier has labelled with the lowest confidence are top-ranked. In this dissertation we show that this strategy is suboptimal. We develop new utility-theoretic ranking methods based on the notion of *validation gain*, defined as the improvement in classification effectiveness that would derive by validating a given automatically labelled document. We also propose new effectiveness measures for SATC-oriented ranking methods, based on the expected reduction in classification error brought about by partially validating a ranked list generated by a given ranking method.

We report the results of experiments showing that, with respect to the baseline method above, and according to the proposed measures, our utility-theoretic ranking methods can achieve substantially higher expected reductions in classification error. We therefore explore the task of SATC and the potential of our methods, in multiple text classification contexts. This dissertation is, to the best of our knowledge, the first to systematically address the task of semi-automated text classification.

Soon available at `http://nmis.isti.cnr.it/berardi`.

Supervisors: Andrea Esuli (ISTI-CNR), Fabrizio Sebastiani (ISTI-CNR).