Text Classification as a Sequential Reading Process: Machine Learning with External Ressources

Ludovic Denoyer (LIP6, Paris 6)

April 8, 2011

We first describe a model that considers text classification as a sequential decision process. In this process, an agent learns to classify documents into topics while reading the document's sentences sequentially. The agent also learns to stop as soon as enough information has been read to properly classify the document. The proposed algorithm is based on a modelisation of Text Classification as a Markov Decision Process and learns by using Reinforcement Learning. We analyze the behaviour of such a process on different classical Text Classification corpora and show that our model seems to adopt an intuitive behaviour: learning to read more sentences when a document is difficult to classify, while being competitive w.r.t. baseline methods.

We then extend this model by authorizing the classification process to access external ressources like dictionnary or wikipedia pages. This correponds to a unique Machine Learning framework where the algorithm not only uses a training set during learning but also external ressources. In a second set of experiments, we show that the process is able to learn when and how to use these external ressources to improve its classification abilities.