Tato práce popisuje systém pro automatické čištění HTML dokumentů, který byl použit při účasti Univerzity Karlovy v soutěži CLEAN-
EVAL 2007. CLEANEVAL je sdílená úloha (shared task) a soutěž automatických systémů pro čištění libovolných stránek s cílem použít webová data jako korpus v počítačové lingvistice a zpracování přirozeného jazyka. Tuto úlohu řešíme jako problém značkování sekvencí (sequence labeling) a náš experimentální systém je založen na algoritmu Conditional Random Fields, používajícím vlastnosti (features) bloků textu odvozené z textového obsahu a HTML struktury analyzovaných webových stránek.