

Streszczenie rozprawy doktorskiej pt.
„Algorithms for automatic grammatical error correction”
w języku polskim

Niniejsza praca doktorska dotyczy problemu automatycznej poprawy błędów językowych w tekstach pisanych przez osoby uczące się języka angielskiego jako języka obcego. Problem ten zbadano za pomocą metod tłumaczenia maszynowego. W celu zgromadzenia dodatkowych danych, opracowano metodę automatycznej ekstrakcji potencjalnych błędów językowych z historii edycji tekstu oraz stworzono największy publicznie dostępny korpus błędów. Zbadano automatyczne miary ewaluacji stosowane w dziedzinie pod kątem ich korelacji z ocenami ludzkimi poprzez wykonanie szczegółowego studium ewaluacji systemów do automatycznej korekty tekstu. Opracowany system, wykorzystujący metody statystycznego tłumaczenia maszynowego opartego na frazach, osiągnął najwyższe publikowane do tej pory wyniki na popularnym zestawie testowym CoNLL-2014 udostępnionym w ramach zadania organizowanego podczas Conference on Natural Language Processing w 2014 roku. W pracy pokazano jak istotne są właściwa optymalizacja modelu na podstawie przyjętej miary ewaluacji oraz zastosowanie nowych cech gęstych. Zbadano również dwie metody integracji algorytmów dyskryminacyjnych do generatywnego systemu tłumaczenia poprzez rozszerzenie modelu log-liniowego. Druga z metod – pierwsze zastosowanie cech rzadkich do zadania korekty tekstu – w sposób istotny rozszerza aktualny stan wiedzy w dziedzinie.

Roman Grundkiewicz