Automatyczne metody wykrywania kłamania.

Prezentacja zawierać będzie omówienie dotychczasowych badań nad automatycznym wykrywaniem fałszywych wypowiedzi w formie tekstowej, ze szczególnym naciskiem na badania w języku polskim. Problem ten obecny jest od kilkunastu lat w eksperymentach z zakresu psycholingwistyki, uczenia maszynowego oraz przetwarzania języka naturalnego (NLP). Badania przeprowadzane w języku angielskim wskazują m.in. na istotną rolę, jaką w wykrywaniu fałszu odgrywają części mowy (part-of-speech) (Newman i in., 2003)W ramach szeroko rozumianej informatyki, problem ten był adresowany na początku obecnego dziesięciolecia w postaci rozpoznawania tzw. opinion spam (Jindal i Liu, 2007; Rubikowski i Wawer, 2013), obecnie zaś pojawia się w kontekście fake news detection (Pszona, Janicka i Wawer 2019, Wawer, Wojdyga i Sarzyńska-Wawer, 2019). W referacie omawiamy dwa główne podejścia do automatycznego rozpoznawania fałszywych treści, jakie pojawiły się na gruncie NLP: psycholingwistyczne i stylometryczne oraz tzw fact-checking, czyli sprawdzanie prawdziwości twierdzeń względem bazy danych tekstów uznawanych za wiarygodne (np Wikipedia). Przedstawimy wady i zalety każdej z metod oraz wstępny zarys najbardziej obiecującej metody hybrydowej, łączącej obydwa podejścia i wykorzystywanej w moich badaniach.

Automated methods of deception detection.

The presentation describes current research on automated detection of false statements in text utterances, with particular emphasis on research in the Polish language. This problem has been present for several years in the field of psycholinguistics, machine learning and natural language processing (NLP). Research conducted in English indicates an important role that part-of-speech plays in detecting deception (Newman et al., 2003) In computer science, this problem was addressed since the beginning of the current decade in the form of recognizing the opinion spam (Jindal & Liu, 2007; Rubikowski & Wawer, 2013), and currently appears in the context of fake news detection (Pszona, Janicka & Wawer 2019, Wawer, Wojdyga & Sarzyńska-Wawer, 2019). In our presentation we discuss two main approaches to automatic recognition of deceptive content in NLP: psycholinguistic and stylometric, and so-called “fact-checking” (checking the truthfulness of statements on the basis of a database of credible texts such as Wikipedia). We will present the pros and cons of each of the methods and an initial outline of the most promising hybrid method combining both approaches.