Nr rejestracyjny: 2018/31/N/ST6/02273; Kierownik projektu: mgr inż. Barbara Rychalska

MAD-NLP: Wieloaspektowa diagnostyka modeli NLP

Dzięki szybkiemu rozwojowi rozwiązań sztucznej inteligencji jest już pewne, że w niedalekiej przyszłości ludzie będą wchodzić w interakcje z maszynami w codziennych sytuacjach. Niedawno firma Google przedstawiła swojego inteligentnego asystenta AI, który działa tak dobrze, że nie sposób odróżnić go od człowieka. Asystent może być używany do interakcji ukierunkowanych na cel, takich jak rezerwacja stolików w restauracji. Chociaż tego typu aplikacje jeszcze nie podejmują samodzielnie żadnej istotnej decyzji, nie można wykluczyć, że w najbliższej przyszłości ich znaczenie będzie rosło.

W tych okolicznościach bardzo ważne jest by stwierdzić jakim poziomem stabilności charakteryzuje się zachowanie modeli. Wiadomo na przykład, że systemy odpowiadające na pytania nie są odporne na zmiany w tekstach wejściowych, co oznacza, że nawet drobne zakłócenia (takie jak wstawienie znaków przestankowych) mogą prowadzić do błędnej odpowiedzi. Możemy postawić hipotezę, że będzie to powodować poważne problemy, szczególnie w krytycznych sektorach takich jak ochrona zdrowia czy dziedziny powiązane z prawem. Przetwarzanie języka naturalnego (ang. *natural language processing*, *NLP*) jest szeroko stosowane w systemach dialogowych (tzw. chatbotach), które pomagają użytkownikom, odpowiadając na ich pytania. Czy w takim razie możemy stwierdzić, że podejmowanie decyzji na podstawie odpowiedzi wirtualnego asystenta jest bezpieczne? Osoby podejmujące poważne decyzje w oparciu o wyniki automatycznego systemu mają prawo, a nawet obowiązek zapytać, na jakich przesłankach opierają się predykcje tych systemów.

Jeszcze bardziej niepokojący jest fakt, że na przewidywania modeli można wpływać celowo, poprzez wstrzykiwanie tzw. *wrogich danych* (ang. *adversarial data*). *Wrogie dane* to przykłady wysyłane do systemu w celu predykcji, które są zmienione w sposób niedostrzegalny dla ludzi, ale które jednocześnie mogą łatwo oszukać głębokie sieci neuronowe. Stworzenie odpowiednio zaprojektowanych *wrogich danych* umożliwia intruzowi sprowokowanie systemu do uzyskania pożądanej odpowiedzi. Może to powodować poważne problemy związane z bezpieczeństwem, na przykład jeśli system błędnie zidentyfikuje znanego terrorystę jako zwykłego obywatela lub przyzna pozytywny ranking kredytowy osobie posiadającej historię oszustw finansowych.

W celu rozwiązania tych problemów proponuję metodologię, która pozwoli zrealizować mój narzędny cel: **Każdy model powinien być testowany pod względem odporności na wrogie przykłady, na równi z klasycznym testowaniem miar jakości modelu**.

Cel 1: Rozwiązanie problemu analizy odporności modeli NLP. Moja metodologia wprowadzi *wieloaspektową* analizę odporności, która zidentyfikuje słabości modelu na wielu płaszczyznach. W przeciwieństwie do obecnego stanu wiedzy, moje metody będą mogły być stosowane do wielu problemów NLP i w sposób całościowy ocenią odporność. Metodologia zostanie przetestowana na wielu problemach NLP, takich jak: odpowiadanie na pytania (QA), analiza sentymentu, tłumaczenie maszynowe (NMT), i inne.

Cel 2: W kolejnym kroku stworzę metodę automatycznej generacji przykładów *wrogich* (ang. adversarial) i *superstabilnych* (ang. *overstable*). Definiuję przykłady wrogie jako teksty, które są zaburzone w minimalny sposób, co jednak zmienia decyzję modelu (choć decyzja powinna pozostać niezmieniona). Natomiast przykłady superstabilne to teksty, które są zaburzone w sposób maksymalny, jednocześnie nie zmieniając decyzji modelu (choć decyzja powinna zostać zmieniona). Moja metoda będzie pozwalała na krokową generację przykładów, co pozwoli na obserwacje postępujących zmian znaczeniowych. Metoda pozwoli na całkowitą eliminację pracy ręcznej przy tworzeniu wrogich przykładów.

Cel 3: Podczas gdy pierwsza i druga faza badań udzieli wglądu w "inteligencję"modelu, faza trzecia wykorzysta metodologię stworzoną w obu poprzednich fazach do analizy związku pomiędzy odpornością a jakością modelu badaną wg. standardowych metryk.

Cel 4: Ostatni cel badań to stworzenie metod korekcji odporności. Metody te będą korzystać z odkryć poczynionych w fazie trzeciej.

Modele, które zastosuję w badaniach odporności (QA, analiza sentymentu i NMT i inne), wykorzystają najnowocześniejsze techniki głębokiego uczenia: rekurencyjne sieci neuronowe, modele enkoderów-dekoderów, konwolucyjne sieci neuronowe. Jednym z założeń mojej pracy jest upowszechnienie stworzonego kodu oraz zbiorów danych, tak by zapewnić weryfikowalność wyników i upowszechnić zastosowanie moich badań w pracach innych członków społeczności NLP.

Podsumowując, mój projekt przyczyni się do rozwoju obszaru NLP, czyniąc algorytmy bardziej zrozumiałymi i bardziej odpornymi. Ich wyniki będą mogły być mierzone w wiarygodny sposób, a ich odporność na ataki z użyciem wrogich danych będzie poprawiona. To zaś przełoży się na komfort i bezpieczeństwo zarówno indywidualnych użytkowników, jak i całych firm.